FDA Briefing Document

Psychopharmacologic Drugs Advisory Committee Meeting

November 4, 2021

Topic: New Drug Application 214812

Carbetocin nasal spray for the treatment of hyperphagia, anxiety, and distress behaviors associated with Prader-Willi syndrome

The attached package contains background information prepared by the Food and Drug Administration (FDA) for the panel members of the advisory committees. The FDA background package contains assessments or conclusions and recommendations written by individual FDA reviewers. Such conclusions and recommendations do not necessarily represent the final positions of the individual reviewers, review Divisions, or Office. FDA has brought New Drug Application 214812 carbetocin nasal spray for the treatment of hyperphagia, anxiety, and distress behaviors associated with Prader-Willi syndrome, to this Advisory Committee in order to gain the Committee's insights and opinions. The background package may not include all issues relevant to the final regulatory recommendation and, instead, is intended to focus on issues identified by FDA for discussion by the Advisory Committee. FDA will not issue a final determination on the issues at hand until input from the Advisory Committee process has been considered and all reviews have been finalized. The final determination may be affected by issues not discussed at the Advisory Committee meeting.

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

GLOSSARY

ANCOVA	analysis of covariance
CARE-PWS	CARbetocin Efficacy and Safety Study in PWS; Study LV-101-3-01
CGI-C	Clinical Global Impression-Change
CGI-I	Clinical Global Impression-Improvement
CGI-S	Clinical Global Impression-Severity
cLDA	constrained longitudinal data analysis
CMC	chemistry, manufacturing, and controls
COVID-19	Coronavirus Disease 2019
CRO	contract research organization
CY-BOCS	Children's Yale-Brown Obsessive-Compulsive Scale
DMC	data monitoring committee
DP	Division of Psychiatry (the primary review division)
DRDMG	Division of Rare Disease and Medical Genetics
EOP2	end-of-phase 2
FAS	full analysis set
FDA	Food and Drug Administration
HPWSQ	Hyperphagia in Prader-Willi Syndrome Questionnaire
HPWSQ-R	Hyperphagia in Prader-Willi Syndrome Questionnaire-Responsiveness
HPWSQ-R-C	Hyperphagia in Prader-Willi Syndrome Questionnaire-Responsiveness-Clinician
HQ-CT	Hyperphagia Questionnaire for Clinical Trials
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
IN	intranasal
IND	investigational new drug
LOCF	last observation carried forward
LTFU	long-term follow-up
MMRM	mixed-model repeated measures
NDA	new drug application
NME	new molecular entity
OCD	obsessive-compulsive disorder
PADQ	PWS Anxiety and Distress Behaviors Questionnaire
PAS	primary analysis set
PDAC	Psychopharmacologic Drugs Advisory Committee
PWS	Prader-Willi syndrome
SAP	statistical analysis plan

1. DIVISION DIRECTOR MEMORANDUM

M E M O R A N D U M

DEPARTMENT OF HEALTH AND HUMAN SERVICES
FOOD AND DRUG ADMINISTRATION
CENTER FOR DRUG EVALUATION AND RESEARCH

DATE: November 4, 2021

FROM: Tiffany R. Farchione, MD
Director
Division of Psychiatry (DP), Office of Neuroscience

TO: Members of the Psychopharmacologic Drugs Advisory Committee (PDAC)

SUBJECT: November 4, 2021, Meeting of the PDAC

Levo Therapeutics, Inc. (the Applicant), has submitted a New Drug Application (NDA) for carbetocin nasal spray (LV-101) for the proposed indication of the treatment of hyperphagia, anxiety, and distress behaviors associated with Prader-Willi Syndrome (PWS). To date, there is no pharmacologic treatment for these aspects of PWS. The application includes data from two studies—a phase 2 proof-of-concept study (Study 114) and a phase 3 study (Study LV 101-3-01). The Applicant believes that the two studies together provide substantial evidence of effectiveness for intranasal (IN) LV-101 3.2 mg three times daily (TID) for the proposed indication; however, the Agency has a number of concerns about the efficacy findings.

Study 114 was a phase 2, randomized, double-blind, placebo-controlled, multicenter, proof-of-concept study of IN carbetocin 9.6 mg TID versus placebo in 38 subjects ages 10 to 18 years with genetically confirmed PWS. Although this study was promising as a proof-of-concept, the statistical results are not robust, the clinical meaningfulness of the observed treatment effect is unclear, and the study duration is insufficient to evaluate a chronic disease for which lifelong treatment is likely needed.

Study LV 101-3-01 was a phase 3, randomized, double-blind, placebo-controlled, multicenter study of TID IN carbetocin 9.6 mg or 3.2 mg versus placebo in subjects ages 7 to 18 years with genetically confirmed PWS, with an 8-week placebo-controlled treatment period followed by a

56-week long-term follow-up (LTFU) period. Although the study planned to enroll 175 subjects, the study was truncated due to the COVID-19 pandemic. A total of 130 randomized subjects received at least one dose of carbetocin. The Applicant set March 1, 2020, as a cutoff date for inclusion in the primary analysis set (PAS); the PAS included 119 subjects who had received at least one dose of carbetocin and completed either the Week 2 or Week 8 visits by the cutoff date.

The two primary endpoints in Study LV 101-3-01 were the change from Baseline to Week 8 for the comparison of IN carbetocin 9.6 mg versus placebo in the Hyperphagia Questionnaire for Clinical Trials (HQ-CT) total score and the Children's Yale-Brown Obsessive-Compulsive Scale (CY-BOCS) Severity Rating total score. Neither of these outcomes were statistically significant. The Applicant suggests that the lack of effect at 8 weeks in the 9.6 mg group may be due to off-target vasopressin effects emerging with longer treatment duration and leading to psychiatric adverse events that may have counteracted the treatment effect. If that were the case, one would expect to see a nominal difference versus placebo at Week 2 (consistent with the findings from Study 114); however, no difference between groups was observed at that earlier time point (p=0.7270).

The first two secondary endpoints were identical to the primary endpoints, but for the comparison of IN carbetocin 3.2 mg versus placebo. Although the *p*-value for the comparison of IN carbetocin 3.2 mg versus placebo for one of two secondary endpoints - the HQ-CT - was nominally significant (p=0.0162), that analysis must be considered in the context of the lack of efficacy of the higher dose (9.6 mg). In addition, results for the second secondary endpoint – the CY-BOCS – showed no meaningful difference between IN carbetocin 3.2 mg and placebo. The lack of efficacy of the 9.6 mg dose questions the biological plausibility of the nominally significant result observed for IN carbetocin 3.2 mg. In addition, results of Study LV-101-3-01 are sensitive to violations of the modeling assumptions, possibly due to the small per arm sample size, further weakening the study results for the 3.2 mg dose of IN carbetocin.

The Applicant also believes that the data from the long-term follow-up period in Study LV 101-3-01 provide support for the efficacy of the 3.2 mg dose of carbetocin. However, given that this portion of the study was open-label, it is subject to confounding by expectation bias, sample size degradation, and potential skewing of results by dropouts for lack of efficacy.

To approve a drug, substantial evidence of effectiveness must be provided by the Applicant. Although two adequate and well-controlled clinical investigations remain the standard for generating substantial evidence of effectiveness in many disease settings, there are scenarios in which a single large multicenter trial can be used to establish effectiveness. As described in the FDA draft guidance on "Demonstrating Substantial Evidence of Effectiveness for Human Drug and Biological Products" (see Attachment), reliance on a single large multicenter trial to establish effectiveness should generally be limited to situations in which the trial has demonstrated a clinically meaningful and statistically very persuasive effect on mortality, severe or irreversible morbidity, or prevention of a disease with potentially serious outcome, and confirmation of the result in a second trial would be impracticable or unethical. Neither of the two studies conducted by the applicant appears to meet these characteristics.

Under certain circumstances, FDA can also conclude that one adequate and well-controlled clinical investigation plus confirmatory evidence is sufficient to establish effectiveness. The aforementioned FDA draft guidance provides examples of the types of confirmatory evidence that can be used, stating:

> FDA will consider a number of factors when determining whether reliance on a single adequate and well-controlled clinical investigation plus confirmatory evidence is appropriate. These factors may include the persuasiveness of the single trial; the robustness of the confirmatory evidence; the seriousness of the disease, particularly where there is an unmet medical need; the size of the patient population; and whether it is ethical and practicable to conduct more than one adequate and well-controlled clinical investigation.

The approach of using one adequate and well-controlled clinical investigation plus confirmatory evidence does not appear applicable to this application, as no confirmatory evidence has been identified.

The purpose of this meeting is to discuss whether substantial evidence of effectiveness for LV-101 in the treatment of hyperphagia associated with PWS has been provided by the Applicant.

DRAFT POINTS TO CONSIDER

The Agency has one question for the Committee:

> Has the Applicant provided substantial evidence of effectiveness for LV-101 in the treatment of hyperphagia associated with Prader Willi syndrome?

2. OBJECTIVE OF MEETING AND OVERVIEW OF DEVELOPMENT PROGRAM

2.1. PURPOSE

The purpose of this Advisory Committee meeting is to discuss NDA 214812 for LV-101 (intranasal (IN) carbetocin), submitted by Levo Therapeutics, Inc. (the Applicant), for the proposed indication of the treatment of hyperphagia, anxiety, and distress behaviors associated with Prader-Willi Syndrome (PWS). The Committee will be asked to discuss whether the Applicant has provided substantial evidence of effectiveness for IN carbetocin in the treatment of hyperphagia associated with PWS.

2.2. PRADER-WILLI SYNDROME

Prader-Willi Syndrome is a rare and serious genetic disorder that manifests with cognitive, behavioral, and endocrine symptoms and physical changes and is associated with a chronic disease course and a shortened life expectancy. Prevalence estimates vary from 1:10,000 to 1:30,000.[1] PWS is caused by the loss of expression of paternally derived genes on chromosome 15q11-13, via deletion, maternal uniparental disomy, or imprinting defects. Phenotypic expression encompasses a wide variety of signs and symptoms, including mild to moderate intellectual disability, learning disabilities, global developmental delay, hyperphagia and a high risk of obesity, emotional and behavioral difficulties, growth hormone deficiency, hypogonadism, and characteristic physical features. Additional symptoms may include scoliosis, seizures, strabismus, sleep abnormalities, and impaired pain sensation, among others.[1, 2]

A longitudinal study of individuals with PWS identified multiple nutritional phases. Infants with PWS exhibit hypotonia, a weak cry and poor suck reflux, and may require tube feeding to avoid failure-to-thrive. This period of poor feeding in infancy is followed by a period of normal growth. However, starting at 20 to 31 months of age, children typically begin to gain weight without a change in appetite or intake. An increased interest in food begins around age 4.5 years, followed by the onset of hyperphagia (with median age of onset at 8 years).[3] Hyperphagia is characterized by a profound lack of satiety, constant preoccupation with food, and food-seeking behavior (e.g., hoarding, foraging, and stealing of food, or eating items such as food from the garbage or pet food).[1,3] In combination with decreased pain sensation and impaired ability to vomit, individuals may binge to the point of acute gastric dilation.[4] Choking episodes have been reported as the cause of death for 6% of individuals with PWS in one sudden death case series.[5] Some adults with PWS may ultimately develop an ability to feel full, but many do not.[3]

Obesity in PWS appears to result from hyperphagia and decreased resting energy expenditure stemming from a combination of decreased lean body mass and decreased activity.[1] The shortened life expectancy in PWS appears to be related mainly to consequences of hyperphagia and obesity, including cardiopulmonary disease and gastrointestinal complications.[6,7]

Management of hyperphagia and obesity consists of a restricted calorie diet, close supervision (including environmental controls such as locking of the refrigerator, cabinets, and trash cans), and exercise. Growth hormone replacement may help by normalizing height, increasing lean body mass, decreasing fat mass, and increasing mobility.[1] No medication has been showed to be effective in treating hyperphagia or obesity, and gastric bypass surgery is not recommended given its lack of benefit for satiety and its complication rates.[1,8] The strict environmental controls that are required to manage hyperphagia significantly impact the quality of life for individuals with PWS and their caregivers, including often limiting the ability to attend social events where food is present. Individuals with PWS may require residence at a specialized group home with strict management of diet, environmental controls, and exercise.

In addition to food-related behavioral difficulties and emotional outbursts, individuals with PWS may experience anxiety, obsessive thoughts and compulsive behaviors, difficulty with change in routine, mood lability, and skin picking, among other symptoms.[1,9] For individuals with PWS caused by maternal uniparental disomy, rates of psychosis are higher than in individuals with other forms of intellectual disability.[9] Selective serotonin reuptake inhibitors have been prescribed off-label for some individuals with PWS with regard to skin picking, compulsivity and aggressive episodes.[10]

2.3. PRODUCT UNDER REVIEW

Intranasal (IN) carbetocin (Applicant code name LV-101) is a synthetic neuroendocrine peptide analogue of the naturally occurring hormone oxytocin. It is a new molecular entity (NME) with the proposed indication of the treatment of hyperphagia, anxiety, and distress behaviors associated with PWS. Although its mechanism of action is unknown, it appears to selectively bind with oxytocin receptors. IN carbetocin is available as an 11.4 mg/mL solution in vials, administered via a nasal spray pump (70 µL per spray); the 3.2 mg dose is delivered as two sprays in each nostril for a total of four sprays per dose, three times daily with meals. Per the Applicant, administration with meals is meant to mimic the natural release of oxytocin concurrent with meals.

2.4. RELEVANT REGULATORY BACKGROUND

- IN carbetocin has not been approved or marketed in the United States. In 1997, carbetocin as a solution for injection was first registered in Canada by the former Sponsor (Ferring Pharmaceuticals, Inc.), for the prevention of uterine atony following caesarean section; carbetocin is approved in a number of countries outside the United States for this indication.

- In June 2012, the Agency and the former Sponsor held a Type C guidance meeting, during which the Agency agreed with the use of the Hyperphagia in PWS Questionnaire (HPWSQ) in an exploratory phase 2 trial. However, the Agency noted that it was unclear that the proposed version of the scale was optimal for a registration trial.

- The Agency suggested assessment of weight loss and improvement in metabolic state as outcomes. The former Sponsor indicated that weight loss and metabolic endpoints may not accurately reflect drug efficacy because of the variable effectiveness of caloric restriction and other environmental interventions, and that the target symptoms are the drive to eat and maladaptive behaviors.

- In August 2013, the former Sponsor submitted the protocol for Study 000114 (Study 114), a phase 2, randomized, double-blind, placebo-controlled, 2-week, proof-of-concept study of IN carbetocin 9.6 mg dose in 38 subjects with PWS aged 10 to 18 years, with a primary endpoint on the revised HPWSQ-Responsiveness (HPWSQ-R) scale. The Agency allowed the study to proceed without comment.

- In April 2014, IN carbetocin received Orphan Drug Designation for the treatment of PWS.

- In February 2015, the Agency denied a Breakthrough Therapy Designation Request based on the results of Study 114. The Agency noted that, although carbetocin showed statistically significant improvement compared to placebo on the HPWSQ-R, the difference between the groups was small.

- In August 2017, the Applicant acquired a license to develop IN carbetocin (referred to as LV-101). The Applicant submitted a letter of authorization to reference IND 112521 in January 2018.

- In May 2018, the Agency held an End-of-Phase-2 (EOP2) pre-IND 138625 meeting with the Applicant to discuss the design of the proposed phase 3 study and the overall development plan.

  - The Applicant inquired whether the Agency agreed that a statistically significant improvement from baseline on either the Hyperphagia Questionnaire for Clinical Trials (HQ-CT) or the Children's Yale-Brown Obsessive-Compulsive (CY-BOCS) scores at 8 weeks in the phase 3 study would be sufficient, together with the Study 114 results, to support NDA submission and review. The Agency disagreed, noting a lack of sufficient justification that an improvement on the CY-BOCS, absent a change on the HQ-CT, would reflect meaningful improvement in a core feature of PWS. The Agency also noted that because PWS will likely require chronic treatment, the Applicant would need to conduct a longer-term safety study in addition to the short-term efficacy studies prior to approval.

  - Following post-meeting submission of information supporting the use of the CY-BOCS in a population of patients with PWS (i.e., outside of its typical use for patients with obsessive-compulsive disorder), the Agency agreed that the CY-BOCS was a potentially appropriate primary endpoint based on the Applicant's hierarchical analysis plan (i.e., with the HQ-CT as the "first" primary and the CY-BOCS as the "second" primary).

- The Agency recommended a thorough dose-finding efficacy study to fully explore the potential dose range and characterize the full dose-response relationship.

- The Agency disagreed with the use of subset analyses of the HQ-CT as secondary endpoints, noting that such measures could only be considered exploratory without sufficient analysis of psychometric properties and objective support for the meaningfulness of such measures, and that removal of items would affect content validity.

- The Agency disagreed with the adequacy of the PWS Anxiety and Distress Behaviors Questionnaire (PADQ) as an observer-reported outcome and requested further evidence to support the content validity of the PADQ. The Division of Clinical Outcome Assessment (DCOA) noted that some items in the PADQ may not be directly observable by caregivers and noted that feelings of anxiety and psychological distress are best known to patients, whereas caregivers or clinicians may only report the observable signs, behaviors, and verbalizations made by patients.

- In September 2018, the Applicant submitted the protocol for Study LV-101-3-01, a phase 3, randomized, double-blind, placebo-controlled, 8-week study of IN carbetocin 9.6 mg or 3.2 mg versus placebo in subjects with PWS ages 7 to 18 years, with a 56-week long-term follow-up, with primary endpoints on both the HQ-CT and CY-BOCS for the 9.6 mg dose, and first secondary endpoints on the HQ-CT and CY-BOCS for the 3.2 mg dose.

- In February 2019, the Agency provided statistical comments, noting that the description of the proposed testing strategy was not clear, and requested clarification regarding whether the Applicant planned to proceed to testing secondary efficacy only when they won on both the primary efficacy endpoints for the 9.6 mg dose. The Agency asked for clarification if achievement of the primary objective was defined as showing statistical significance on *either* of the primary efficacy endpoints. In addition, the Agency stated that the proposed constrained longitudinal data analysis (cLDA) may be acceptable and similar to MMRM. The acceptability of cLDA was determined to be a review issue in this communication. The Applicant never responded to the Agency to clarify their multiple comparison procedure (MCP).

- In February 2019, Fast Track Designation was granted to LV-101 for PWS.

- In April 2020, the Applicant submitted a Letter of Amendment for Study LV-101-3-01 in response to the Coronavirus Disease 2019 (COVID-19) pandemic. Because of the potential for pandemic-related symptom changes to confound outcomes, the Applicant proposed to include in a Primary Analysis Set (PAS) only subjects who completed the Baseline visit and one post-Baseline visit (i.e., Week 2 or Week 8) prior to March 1, 2020. The Agency did not object to the proposal.

- At a November 2020 pre-NDA meeting:
  - The Agency disagreed that Study LV-101-3-01, together with the results of Study 114, would support the submission and filing of an NDA.
  - The Agency noted that Study LV-101-3-01 failed on both primary endpoints for the 9.6 mg dose, and that based on the prespecified statistical testing hierarchy, testing should have stopped there.
  - As also noted at the May 2018 EOP2 meeting, the Agency disagreed with the secondary endpoints of the HQ-CT subsets and the PADQ.
  - The Agency reiterated its previous assessment that the results of Study 114 were of unclear clinical significance and the study was too short to support a chronic indication.
  - The Agency acknowledged the Applicant's hypothesis regarding a U-shaped dose-response curve but noted that such curves are rare, and the results of the phase 2 and 3 studies were inconsistent.
  - The Agency noted that in the cases of rare diseases, the Agency can be flexible with how many patient exposures are acceptable, but the Applicant still needs to meet the statutory requirement to demonstrate substantial evidence of effectiveness. The Agency recommended that the Applicant test the 3.2 mg dose in another study.
  - In a post-meeting comment from the Division of Rare Disease and Medical Genetics (DRDMG), DRDMG agreed with the Division of Psychiatry that the results of the clinical studies of LV-101 did not appear to be adequate to support an NDA. DRDMG noted that the rare disease status of PWS does not exempt the program from having to meet the statutory requirement to demonstrate substantial evidence of effectiveness.
- The Applicant submitted the NDA on April 13, 2021. After consultation with senior leadership, it was decided that the application could be filed, with questions regarding the study designs and the statistical and clinical meaningfulness of the findings to be addressed in a comprehensive review of the Applicant's data.
- The Applicant did not submit the Statistical Analysis Plan (SAP) for either Study 114 or Study LV-101-3-01 to the Agency for review during the IND. Study LV-101-3-01's SAP was first submitted to the Agency with the NDA on April 13, 2021, after the database had been locked and the analysis had been completed. The SAPs were not reviewed by Agency before database lock; therefore, no feedback from the Agency was incorporated.
- The application was filed on June 11, 2021. Because the product under review received a Fast Track designation and it is intended to address an unmet medical need, the application was granted a priority review designation.

3. SUMMARY OF CLINICAL AND STATISTICAL DATA

3.1. EFFECTIVENESS OF INTRANASAL CARBETOCIN

The Applicant submitted the results of two studies to support the efficacy and safety of IN carbetocin. Study 114 was a phase 2, 2-week, proof-of-concept study. Study LV-101-3-01 was a phase 3 study that included an 8-week placebo-controlled period and a 56-week long-term follow-up period.

3.1.1. Study 000114

Study 114 was titled "A Prospective, Randomized, Double-Blind, Placebo-Controlled Study to Evaluate the Safety and Effectiveness of Intranasal Carbetocin in Subjects with Prader-Willi Syndrome (PWS)."

Study Design

Study 114 was a phase 2, randomized, double-blind, placebo-controlled, multicenter, proof-of-concept study of IN carbetocin 9.6 mg versus placebo in subjects ages 10 to 18 years with genetically confirmed PWS. Subjects were randomized 1:1 for a 14-day placebo-controlled treatment period, with final site visit on Day 15 and a telephone follow-up contact on Day 19.

Dosing

Subjects were to receive study drug three times daily with meals. No dose modification was available.

Study Schedule

Study visits occurred on Days 1, 2, and 15 on-site, with off-site assessment on Day 8 and a telephone follow-up on Day 19. See Table 7 in the Appendix for the Applicant's detailed schedule of assessments.

Study Endpoints

The primary endpoint was the change from Baseline (Day 1) to End-of-Treatment (Visit 4, Day 15) in the HPWSQ-R total score.

The HPWSQ-R (a revised version of the HPWSQ) is an 11-item observer-reported outcome measure designed to be used for repeated measures to evaluate the change in hyperphagia severity after intervention over a 1-week recall period. The HPWSQ-R is subdivided into three domains, each with their own subscore. Items are rated on a 5-point ordinal scale, where 1 indicates the behavior has not been observed and 5 indicates the behavior was very severe or frequent (except for ease of redirection, rated extremely easy to extremely hard). Total scores range from 11 to 55 points; higher scores indicate more severe or frequent behaviors, and negative change indicates improvement. See the Appendix for a copy of the HPWSQ-R.

Secondary endpoints included: Clinical Global Impression-Improvement (CGI-I) score at End-of-Treatment; change from Baseline to End-of-Treatment for the HPWSQ-R hyperphagia, drive, and severity domain scores, the HPWSQ-R-Clinician (HPWSQ-R-C) total score, and the HPWSQ-R-C hyperphagia, drive, and severity domain scores; and change from Screening (Visit 1) to End-of-Treatment for the CY-BOCS score and the Food Domain of the Reiss Profile.

The CGI-I is clinician-rated scale that assesses how much an individual's overall condition has changed from baseline.

The CY-BOCS is a clinician-administered, semi-structured inventory of obsessions and compulsions with a clinician-rated measure of symptom severity. The CY-BOCS consists of a symptom checklist for current and past obsessions and compulsions and a 10-item severity rating scale for obsessions and compulsions (five items each), including time occupied by, interference due to, distress associated with, resistance against, and degree of control over obsessions or compulsions. Each item is rated from 0 (none) to 4 (extreme) for a maximum score of 40 points; higher scores indicate more severe or frequent symptoms, and negative change indicates improvement. The maximum score is 40 points; higher scores indicate more severe or frequent symptoms, and negative change indicates improvement. See the Appendix for a copy of the CY-BOCS.

The Food Domain of the Reiss Profile consists of seven items that pertain to food-seeking behavior, rated on a 5-point scale ranging from "strongly disagree" (-2) to "strongly agree" (2).

Statistical Considerations

All statistical tests were performed by the Applicant using a one-sided test at a 0.1 significance threshold (compared to the one-sided 0.025 level that is typical for an adequate and well-controlled study). The last observation carried forward (LOCF) method was used to carry forward non-missing values of HPWSQ-R total score during the Day 8 phone call to impute missing values of HPWSQ-R total score at End-of-Treatment (Day 15), as well as for the HPWSQ-R-C. Data were analyzed in the full analysis set (FAS) population using an analysis of covariance (ANCOVA) model with treatment and site as fixed effects and HPWSQ-R total score at baseline as a covariate. There was no pre-specified plan to control type I error rate for multiple comparisons over the primary and secondary endpoints. The sample size, power, and alpha level were adequate for a proof-of-concept trial.

Study Results

Thirty-eight subjects were screened and randomized. One subject in the IN carbetocin arm was a screening failure randomized in error and did not receive any study drug, so the FAS and safety analysis set both included 37 subjects: 17 subjects in the IN carbetocin 9.6 mg arm and 20 subjects in the placebo arm.

Disposition

Of the 37 subjects dosed, one subject in the IN carbetocin arm discontinued because of adverse events (AEs) including agitation, increased aggression, increased hyperphagia, and broken distal ulna (see Section 3.2). All other subjects completed the trial.

Demographics and Baseline Characteristics

The mean age of subjects was 13.7 years. Subjects included 62% females and 38% males. Most subjects across treatment arms identified as White (approximately 97%) and Not Hispanic or Latino (approximately 97%). Mean body weight was 63.9 kg and mean body mass index was 25.7 kg/m^2. Most demographic and baseline characteristics were relatively similar between treatment groups, except the IN carbetocin group had a somewhat larger mean baseline weight (66.3 kg versus 61.8 kg for placebo). Mean baseline HPWSQ-R scores were slightly more severe for the placebo group (39.7 versus 35.6 for IN carbetocin). Mean baseline CY-BOCS scores were similar in the drug and placebo groups (approximately 15.8). Overall, the demographic and baseline characteristics of the treatment arms were similar enough to permit meaningful comparison on the study endpoints.

Efficacy Results – Primary Endpoint

As originally reported, on the primary endpoint of HPWSQ-R total score change from Baseline to Day 15, IN carbetocin 9.6 mg separated from placebo at a one-sided p-value of 0.0290 (two-sided 0.0580, see Table 1). During an FDA inspection of a study site, an Agency investigator noted a discrepancy between the paper source and the study database for one subject on the HPWSQ-R Day 15 results; the correction resulted in a one-sided p-value of 0.0244 (two-sided 0.0488, see Table 2). Although the corrected results were statistically significant at both the prespecified 0.10 threshold and the 0.025 threshold more typical of confirmatory trials, the clinical meaningfulness of the results was unclear, as a 2-week endpoint is insufficient to assess potential efficacy in a chronic condition. Further data from the longer phase 3 study would be required to better understand IN carbetocin's effects.

Table 1: Study 000114 Primary Endpoint Efficacy Results (Full Analysis Set)

Primary endpoint HPWSQ-R total score	IN carbetocin 9.6 mg (N=17)	Placebo (N=20)
Baseline mean (SD)	35.6 (7.20)	39.7 (7.62)
Change from baseline to Day 15		
LS mean (SE)	-15.6 (3.06)	-8.9 (2.61)
LS mean difference versus placebo	-6.7	
Upper limit of 90% CI	-2.2	
95% CI	-13.2, 0.10	
One-sided p-value	0.0290	
Two-sided p-value	0.0580	

Source: Adapted from Study 000114 Clinical Study Report, Table 9-1; verified by Statistical Reviewer
CI = confidence interval, LS = least squares, SD = standard deviation, SE = standard error
Note: For placebo Subject (b) (6), the Day 8 assessment was imputed for invalid values at Day 15, because it was an early termination visit after discontinuation for an adverse event.

Table 2: Study 000114 Primary Endpoint Efficacy Results (Full Analysis Set, Corrected Results)

Primary endpoint HPWSQ-R total score	IN carbetocin 9.6 mg (N=17)	Placebo (N=20)
Baseline mean (SD)	35.6 (7.20)	39.7 (7.62)
Change from baseline to Day 15		
LS mean (SE)	-15.7 (3.05)	-8.8 (2.60)
LS mean difference versus placebo	-6.9	
Upper limit of 90% CI	-2.5	
95% CI	-13.4, -0.11	
One-sided p-value	0.0244	
Two-sided p-value	0.0488	

Source: Adapted from Clinical Information Amendment, revised Table 14.2-1; verified by Statistical Reviewer
CI = confidence interval, LS = least squares, SD = standard deviation, SE = standard error
Note: For placebo Subject (b) (6), the Day 8 assessment was imputed for invalid values at Day 15, because it was an early termination visit after discontinuation for an adverse event.

Efficacy Results – Secondary Endpoints

Among secondary endpoints of note, the CGI-I score at Day 15 resulted in an estimated effect of -0.8 (90% CI upper limit: -0.3; 95% CI: -1.62 to -0.12) for IN carbetocin compared with placebo (one-sided p-value of 0.0233), and the CY-BOCS total score change from baseline to Day 15 resulted in an estimated effect of -6.2 (90% CI upper limit: -3.3; 95% CI: -10.4 to -1.67) for IN carbetocin compared with placebo (one-sided p-value of 0.0047). Secondary endpoints were not controlled for multiple comparisons under a pre-specified plan.

Robustness of Results

Of the 38 subjects randomized, six subjects (16%) were considered to have major protocol violations, including a screen failure subject who was randomized in error to IN carbetocin (but not dosed) and a placebo subject who discontinued for adverse events (who was considered to have substandard dosing compliance according to the SAP). Two other placebo subjects had major violations, including two subjects who were erroneously shipped the wrong kit (both on Day 5) and dosed from the incorrect bottle. Two IN carbetocin subjects had violations: one had the HPWSQ-R completed on Day 16 instead of 15, and data were not able to be used; the other was erroneously shipped the wrong kit (on Day 4) and dosed three times from the incorrect bottle. (A fourth subject was erroneously shipped the wrong kit but did not dose from the incorrect bottle and was considered a minor violation.) As demonstrated by the change in the primary endpoint result with the corrected data for one subject above, issues with study conduct in a small proof-of-concept study such as this can have an impact on the reliability of study results.

3.1.2. Study LV-101-3-01

Study LV-101-3-01 was titled "Phase 3, Randomized, Double-Blind, Placebo-Controlled, 8-Week Clinical Study to Assess the Efficacy, Safety, and Tolerability of Intranasal Carbetocin (LV-101) in Prader-Willi Syndrome (PWS) With Long-Term Follow-Up: CARE-PWS."

Study Design

Study LV-101-3-01 was a phase 3, randomized, double-blind, placebo-controlled, multicenter study of IN carbetocin 9.6 mg or 3.2 mg versus placebo in subjects ages 7 to 18 years with genetically confirmed PWS, with an 8-week placebo-controlled treatment period followed by a 56-week long-term follow-up (LTFU) period. The Applicant originally planned to enroll 175 patients. For inclusion, subjects were required to have a baseline HQ-CT score ≥ 13 and a baseline CY-BOCS score ≥ 9. No new food-related interventions, including environmental or dietary restrictions, were permitted within 1 month of Screening, and doses of any permitted chronic concomitant medications were to be stable for ≥ 3 months prior to and during the study. Subjects were randomized 1:1:1 for the 8-week placebo-controlled treatment period. Subjects randomized to placebo for the 8-week period were further randomized to receive IN carbetocin 9.6 mg or 3.2 mg for the 56-week LTFU period, whereas subjects initially randomized to IN carbetocin continued the same dose they were receiving in the double-blind period. The dose of IN carbetocin remained blinded to investigators, caregivers, and subjects throughout the LTFU period. Following completion of the LTFU period, subjects could continue in an ongoing open-label extension.

Dosing

Subjects were to receive study drug three times daily with meals. No dose modification was permitted during the 8-week placebo-controlled treatment period. Following completion of the 8-week period and data review and recommendations by the Data Monitoring Committee (DMC), the Applicant could allow subjects to continue on their assigned dose in the LTFU and extension or switch them to a dose they deemed safer or more efficacious.

Study Schedule

For the 8-week placebo-controlled treatment period, study visits occurred at Screening, Baseline, Week 2, and Week 8. For the LTFU period, study visits occurred at Weeks 10, 16, 28, 40, 52, and 64. See the Appendix for the Applicant's detailed schedule of assessments for both periods.

Study Endpoints

The two primary endpoints were the change from Baseline to Week 8, for the comparison of IN carbetocin 9.6 mg versus placebo, in the HQ-CT total score and the CY-BOCS Severity Rating total score. The first two secondary endpoints were identical to the primary endpoints (i.e., change from Baseline to Week 8 in the HQ-CT total score and the CY-BOCS Severity Rating total score), but for the comparison of IN carbetocin 3.2 mg versus placebo. As noted, following post-May 2018 meeting submission of information supporting the use of the CY-BOCS in a population of patients with PWS (i.e., outside of its typical use for patients with obsessive-compulsive disorder), the Agency agreed that the CY-BOCS was a potentially appropriate primary endpoint based on the Applicant's then-proposed hierarchical analysis plan (i.e., with the HQ-CT as the "first" primary and the CY-BOCS as the "second" primary). The Applicant did not respond to the Agency's February 2019 statistical comments requesting clarification of their then-proposed statistical testing plan and MCP.

Other secondary endpoints, tested first for IN carbetocin 9.6 mg versus placebo then for IN carbetocin 3.2 mg versus placebo, included change from Baseline to Week 8 on the PADQ score, the Clinical Global Impression-Change (CGI-C) score at Week 8, change from Baseline to Week 8 on an HQ-CT subset (items 1, 2, 5, 6, 8, 9), and change from Baseline to Week 8 on HQ-CT item 9.

The HQ-CT (a revised version of the HPWSQ-R) is a 9-item outcome measure designed to assess the frequency and severity of caregiver-observed behaviors associated with hyperphagia among subjects with PWS in clinical trials over a 2-week recall period, including such behaviors as being upset when denied food, bargaining for more food, or time talking about food. Items are rated on a 5-point ordinal scale, where 0 indicates the behavior has not been observed and 4 indicates the behavior was very severe or frequent. The items in the HQ-CT are similar to those in the HPWSQ-R, with the two items from the HPWSQ-R removed (how "clever" or "fast" the patient is in obtaining food and how easy it is to redirect the patient away from food). Total scores range from 0 to 36 points; higher scores indicate more severe or frequent behaviors, and negative change indicates improvement. See the Appendix for a copy of the HQ-CT.

According to the Applicant, the PADQ is a "caregiver-reported instrument designed to capture the observable signs of anxiety and distress that are common among subjects with PWS." The PADQ is 15-item questionnaire, with items rated from 0 (never) to 4 (always or almost always). The first 14 items are scored for a maximum score of 56 points; higher scores indicate more frequent symptoms, and negative change indicates improvement. Item 15 asks about an overall summary of caregivers' impressions but is not included in the total score. As noted, the Agency disagreed with the adequacy of the PADQ as an observer-reported outcome and requested further evidence to support the content validity of the PADQ. The Division of Clinical Outcome Assessment noted that some items in the PADQ may not be directly observable by caregivers and noted that feelings of anxiety and psychological distress are best known to patients, whereas caregivers or clinicians may only report the observable signs, behaviors, and verbalizations made by patients.

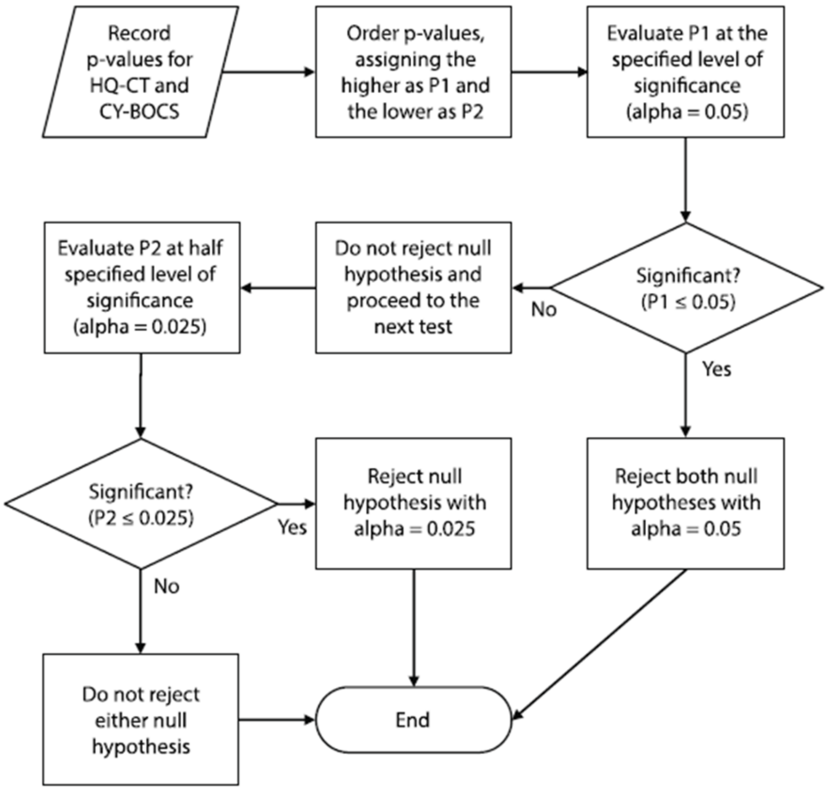
The CGI-C was identical to the CGI-I used in Study 114.

The HQ-CT subset that was included as a secondary endpoint eliminated the items (3, 4, and 7) that the Applicant considered susceptible to environmental controls and routines to restrict access to food. If food is not accessible, the hyperphagia-related behaviors assessed by these items—getting up at night to food seek, foraging through trash, or trying to sneak or steal food—may be infrequent. The Applicant also asserted that HQ-CT item 9, which asks the caregiver to rate food-related behavior interference with normal daily activities, is valuable on its own in assessing the burden of the condition. As previously stated, the Agency disagreed with the use of subset analyses of the HQ-CT as secondary endpoints, noting that such measures could only be considered exploratory without sufficient analysis of psychometric properties and objective support for the meaningfulness of such measures, and that removal of items would affect content validity.

Statistical Considerations

Overall Type I error was specified at the two-sided 0.05 level of significance. The Applicant proposed that 175 patients would provide 90% for the HQ-CT endpoint and 99% power for the CY-BOCS endpoint assuming that the phase 2 results (HQ-CT: 5.5 mean difference with standard deviation = 8.5; CY-BOCS: 5.7 mean difference with standard deviation = 6.4) held in the phase 3 study. The Hochberg step-up procedure was used to control for multiplicity across the two primary endpoints (HQ-CT and CY-BOCS change from baseline to Week 8 for the 9.6 mg dose): if the larger p-value was ≤ 0.05, both p-values were considered significant; if the larger p-value was > 0.05, then the smaller p-value would be compared to the ≤ 0.025 threshold for significance (Figure 1). The Applicant's SAP stated that the secondary endpoints would be tested if at least one of the two tests on the 9.6 mg dose were statistically significant: if both primary hypothesis tests were rejected, alpha of 0.05 would be passed to the secondary endpoints; otherwise, if one of two primary hypothesis tests were rejected, alpha of 0.025 would be passed to the secondary endpoints. Type I error was proposed to be controlled for the first two secondary endpoints (HQ-CT and CY-BOCS change from baseline to Week 8 for the 3.2 mg dose) using the same Hochberg step-up procedure as the primary endpoint. For the additional secondary endpoints (PADQ, CGI-C, HQ-CT subset, and HQ-CT item 9), the Applicant proposed to control type I error using a hierarchical method with the alpha passed on from the first two Hochberg procedures.

Figure 1: Flow Chart for Hochberg's procedure used in Study LV-101-3-01



Source: Applicant's SAP for Study LV-101-3-01

As noted in Section 2.4 (Relevant Regulatory Background), the Applicant did not submit the SAP to the FDA for review during the IND. The SAP was first submitted to FDA with the NDA. The Applicant's multiple testing procedure does not control the overall Type I error rate. To control the overall Type I error rate in this case, no alpha should be passed to the secondary endpoints unless both p-values from the primary endpoints were significant; that is, if at least one of the two p-values is not significant, testing should be stopped, and subsequent statistical evaluations should be descriptive.

The Applicant pre-specified a constrained longitudinal data analysis (cLDA) model for the primary analysis model. The cLDA model has an outcome vector that includes the baseline value as an outcome instead of a covariate. In addition, the cLDA model included treatment arm, visit (Baseline, Week 4, Week 8), and site as fixed effects with an assumption of compound symmetry. A further assumption was made that the three treatment arms had the same baseline value. Finally, the Applicant assumed that missing data was missing at random.

As noted in Section 2.4 (Relevant Regulatory Background), the Applicant did not submit the SAP to the FDA for review during the IND. The Applicant's assumption of a compound symmetry covariance structure may lead to an underestimate of the standard errors. The standard practice in phase 3 clinical trials is to use an unstructured covariance matrix unless the model does not converge.

Furthermore, the cLDA model assumed that all treatment arms have the same means at baseline. For a large, randomized trial, the assumption that all treatment arms will be balanced on the baseline outcome value will be satisfied. However, there may be baseline imbalance for small trials, which may bias the treatment effect estimates.

Study Results

Because of the COVID-19 pandemic, the Applicant temporarily held new screening and enrollment on March 12, 2020, subsequently closed further screening and enrollment on May 30, 2020, and ended follow-up for efficacy on July 31, 2020. The last subject was enrolled on February 26, 2020. The Applicant defined two analysis sets:

- The primary analysis set (PAS) consisted of 119 randomized subjects who received at least one dose of carbetocin and completed either the Week 2 or Week 8 visits before March 1, 2020.

- The full analysis set (FAS) consisted of 130 randomized subjects who received at least one dose of carbetocin.

All analyses were conducted on the PAS with sensitivity analyses conducted on the FAS. Full details of the analysis sets are in Table 3.

Table 3: Study LV-101-3-01 Analysis Sets (All Randomized Subjects)

	IN carbetocin 3.2 mg (N=43) n (%)	IN carbetocin 9.6 mg (N=44) n (%)	Placebo (N=43) n (%)	Total (N=138) n (%)
Safety Analysis Set[a]	43 (100.0)	44 (100.0)	43 (100.0)	130 (94.2)
Full Analysis Set[b]	43 (100.0)	44 (100.0)	43 (100.0)	130 (94.2)
Primary Analysis Set[c]	39 (90.7)	40 (90.9)	40 (93.0)	119 (86.2)
Per-Protocol Analysis Set[d]	36 (83.7)	33 (75.0)	36 (83.7)	105 (76.1)

Source: Adapted from Study LV-101-3-01 Clinical Study Report, Table 12
[a] The Safety Analysis Set included all subjects who received at last one dose of investigational product. Counts are based on treatment received.
[b] The Full Analysis Set included subjects who were both randomized and dosed. Counts are based on randomized treatment.
[c] The Primary Analysis Set included all subjects in the FAS with at least one post-Baseline visit completed prior to March 1, 2020. Counts are based on randomized treatment.
[d] The Per-Protocol Analysis Set included subjects in the PAS who did not meet criteria for PPS exclusion as outline in the Statistical Analysis Plan. Counts are based on randomized treatment.

Disposition

The 130 subjects in the FAS included 44 subjects in the 9.6 mg arm and 43 subjects each in the 3.2 mg and placebo arms. Of those, 128 subjects (98.5%) completed the Week 8 visit and entered the LTFU period. Two subjects (4.5%) in the 9.6 mg arm discontinued during the placebo-controlled period for AEs (impulsive behavior and tachycardia).

For the LTFU period, of the 43 subjects in the placebo arm, 22 subjects were randomized to 9.6 mg and 21 subjects to 3.2 mg. Subjects on IN carbetocin during the placebo-controlled period stayed on the same dose for the LTFU period. In total, the 9.6 mg and 3.2 mg arms entered the LTFU period with 64 subjects each. As of the 120-day safety update, 25 subjects discontinued during the LTFU period, including 17 subjects in the 9.6 mg arm and eight subjects in the 3.2 mg arm.

Demographics and Baseline Characteristics

Mean age was approximately 12 years in all treatment arms. The 9.6 mg and placebo arms each included a relatively even distribution of male and female subjects; the 3.2 mg arm included somewhat more females than males (62% versus 39%). Most subjects across treatment arms identified as White (approximately 85%) and Not Hispanic or Latino (approximately 90%). The most frequent PWS genetic subtype was deletion type unknown, followed by uniparental disomy. The placebo arm had a lower baseline mean weight and body mass index (53.6 kg and 24.3 kg/m^2, respectively) compared to the all-dose IN carbetocin group (60.2 kg and 26.4 kg/m^2, respectively).

In Table 4, the baseline summaries of the primary and secondary endpoints are presented. The mean HQ-CT baseline score was higher in the 9.6 mg arm, with a baseline of 23.4 points, compared to 22.1 points for the 3.2 mg arm, and 22.4 points for the placebo arm. Figure 2 showed boxplots (the boxes are the 25th to 75th percentiles with the median indicated as a solid line and the mean value indicated with a solid dot) of the HQ-CT at each visit. In this figure, the 9.6 mg arm has a higher median and mean HQ-CT at baseline, with a higher variation compared to placebo and the 3.2 mg arm. This trend continues at Week 2 and Week 8. Furthermore, the 9.6 mg arm and the placebo arm had more subjects with a baseline HQ-CT score of at least 30 compared to 3.2 mg arm (see Figure 3). This figure displays the individual response trajectories from the baseline visit to Week 8. In both the 9.6 mg arm and the placebo arm, there were more lines starting above 30 points at the baseline visit than in the 3.2 mg arm. The mean CY-BOCS baseline score was lower in the 3.2 mg arm, with a mean of 25.2 points, compared to 28 points in the 9.6 mg and placebo arms. The mean PADQ baseline score was lowest in the 9.6 mg arm.

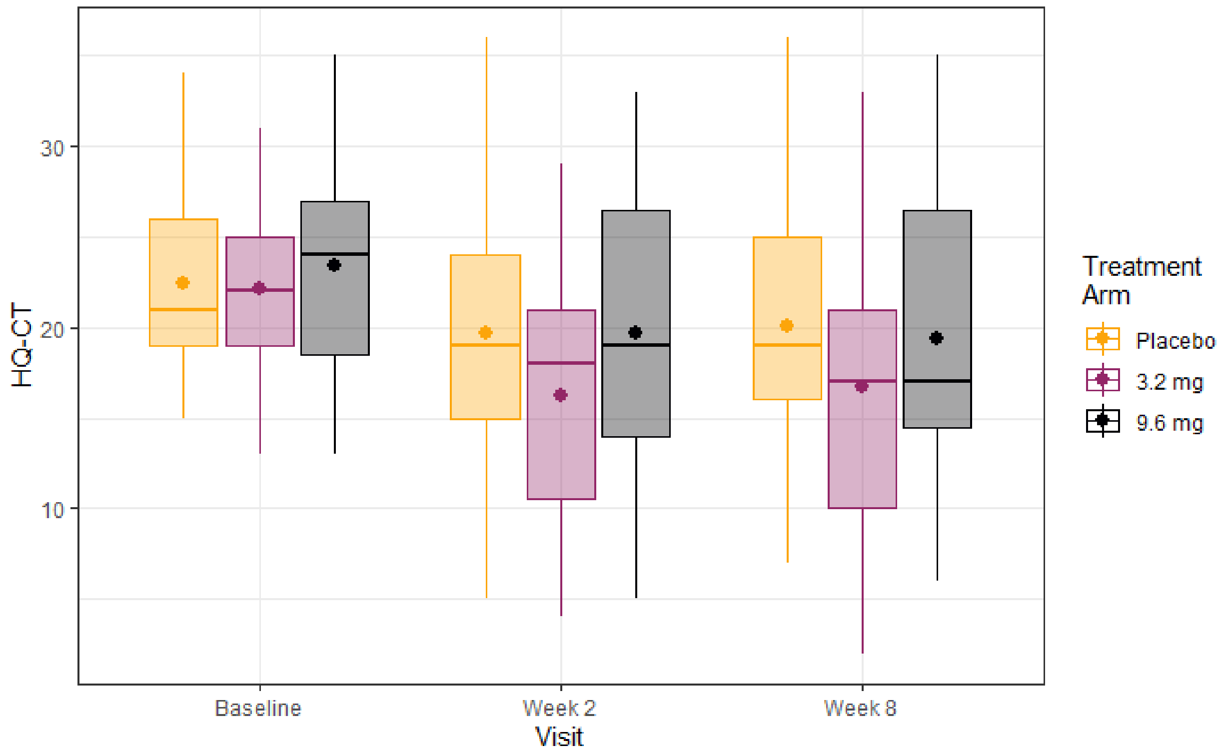
Table 4: Study LV-101-3-01 Baseline Primary Endpoint (Primary Analysis Set)

	IN carbetocin 3.2 mg (N=39)	IN carbetocin 9.6 mg (N=40)	Placebo (N=40)	Total (N=119)
CY-BOCS				
Mean (SD)	25.2 (4.5)	27.9 (5.1)	27.8 (6.0)	27.0 (5.3)
Median	26.0	28.5	27.5	27.0
Min, Max	13, 33	8, 35	15, 40	8, 40
25th percentile	23.0	26.0	25.0	24.0
75th percentile	28.0	31.0	32.2	30.5
HQ-CT				
Mean (SD)	22.1 (5.1)	23.4 (5.7)	22.4 (4.7)	22.6 (5.2)
Median	22.0	24.0	21.0	22.0
Min, Max	13, 31	13, 35	15, 34	13, 35
25th percentile	19.0	18.5	19.0	19.0
75th percentile	25.0	27.0	26.0	26.0
PADQ				
Mean (SD)	43.1 (6.9)	42.5 (7.2)	43.9 (6.7)	43.2 (6.9)
Median	43.0	44.5	44.0	44.0
Min, Max	24, 56	29, 55	32, 55	24, 56
25th percentile	38.5	36.0	39.8	37.0
75th percentile	47.5	48.2	49.5	48.0

Source: Statistical Reviewer Analysis; adsl.xpt

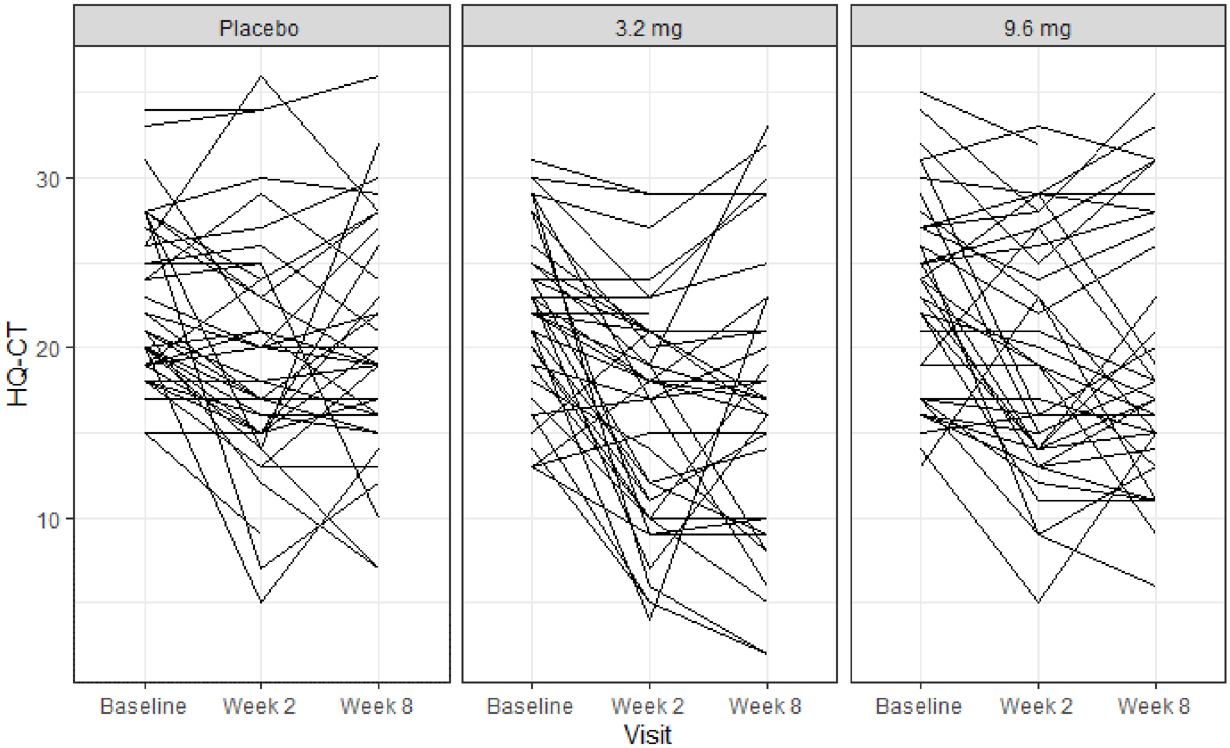
CY-BOCS = Children’s Yale-Brown Obsessive-Compulsive Scale, HQ-CT = Hyperphagia Questionnaire for Clinical Trials, Min = minimum, Max = maximum, PADQ = PWS Anxiety and Distress Behaviors Questionnaire, SD = standard deviation

Figure 2: Study LV-101-3-01 HQ-CT Box Plots at Each Visit (Primary Analysis Set)



Source: Statistical Reviewer Analysis; adef.xpt
HQ-CT = Hyperphagia Questionnaire for Clinical Trials

Figure 3: Study LV-101-3-01 Subject HQ-CT Trajectories (Primary Analysis Set)



Source: Statistical Reviewer Analysis; adef.xpt
HQ-CT = Hyperphagia Questionnaire for Clinical Trials

<u>Efficacy Results – Primary Endpoints</u>

For both primary endpoints (HQ-CT and CY-BOCS change from Baseline to Week 8), Study LV-101-3-01 did not show a significant improvement in the 9.6 mg carbetocin arm compared to placebo (Table 5). For the HQ-CT endpoint, the least squares (LS) mean difference versus placebo for the carbetocin 9.6 mg arm was -1.2 points (95% CI -3.7 to 1.3; *p*-value = 0.3493). For the CY-BOCS endpoint, the treatment difference between carbetocin 9.6 mg and placebo was -0.608 (95% CI: -2.89 to 1.67; *p*-value = 0.6001). At this point, statistical testing was to stop per the pre-specified multiple comparison procedure because both *p*-values are greater than 0.05. Therefore, Study LV-101-3-01 did not meet its pre-specified primary endpoint. In addition, neither endpoint at the 9.6 mg dose arm showed any improvement compared to placebo at Week 2 (Table 5). This finding is inconsistent with the findings of Study 114, in which 9.6 mg dose arm appeared to show a modest treatment effect at 15 days. Therefore, Study LV-101-3-01 failed to confirm the results of the phase 2 program.

Table 5: Study LV-101-3-01 Efficacy Results (Primary Analysis Set)

Endpoint	Visit		IN carbetocin 9.6 mg (N = 40) – Primary	IN carbetocin 3.2 mg/dose (N = 39) – Secondary	Placebo (N = 40)
HQ-CT	Baseline	Mean[a] (SE)		22.22 (0.730)	
	Week 2	LS mean CFB (SE)	-3.402 (0.905)	-5.983 (0.907)	-2.977 (0.894)
		LS mean difference vs. placebo	-0.4250	-3.006	--
		(95% CI of LS mean differences)	(-2.821, 1.971)	(-5.410, -0.602)	--
		Two-sided p-value	0.7270	0.0145	--
	Week 8	LS mean CFB[b] (SE)	-3.439 (0.946)	-5.372 (0.957)	-2.237 (0.943)
		LS mean difference vs. placebo	-1.202	-3.136	--
		(95% CI of LS mean differences)	(-3.729, 1.324)	(-5.685, -0.586)	--
		Two-sided p-value	0.3493	0.0162	--
CY-BOCS	Baseline	Mean (SE)		27.33 (0.632)	
	Week 2	LS mean CFB (SE)	-2.571 (0.824)	-3.037 (0.826)	-1.220 (0.810)
		LS mean difference vs. placebo	-1.351	-1.818	--
		(95% CI of LS mean differences)	(-3.514, 0.811)	(-3.990, 0.353)	--
		Two-sided p-value	0.219	0.1003	--
	Week 8	LS mean CFB (SE)	-2.968 (0.862)	-3.123 (0.873)	-2.360 (0.855)
		LS mean difference vs. placebo	-0.608	-0.764	--
		(95% CI of LS mean differences)	(-2.890, 1.674)	(-3.068, 1.541)	--
		Two-sided p-value	0.6001	0.5143	--

Source: Modified from Applicant's Summary of Clinical Efficacy by the Statistical Reviewer; adef.xpt
CFB = change from baseline, CI = confidence interval, CY-BOCS = Children's Yale-Brown Obsessive-Compulsive Scale, HQ-CT = Hyperphagia Questionnaire for Clinical Trials, LS = least squares, SE = standard error
[a] Baseline mean is estimated from the cLDA primary analysis model

Efficacy Results – Secondary Endpoints

On the HQ-CT change from Baseline to Week 8, IN carbetocin 3.2 mg showed a -3.14 (95% CI -5.69, -0.59) LS mean difference from placebo. On the CY-BOCS, IN carbetocin 3.2 mg did not separate from placebo with a treatment difference of -0.76 (95% CI: -3.07 to 1.54). Because statistical testing stopped after the testing the two 9.6 mg endpoints, the p-values reported in Table 5 are nominal and must be considered as descriptive statistics only.

For the additional secondary endpoints in the 9.6 mg dose arm, none of them showed any signal of improvement compared to placebo. For the 3.2 mg dose, the PADQ total score change from Baseline to Week 8 and the CGI-C had descriptive *p*-values of 0.027, while the HQ-CT subset and item 9 had descriptive *p*-values of 0.011 and 0.114, respectively. However, these *p*-values cannot be used to draw statistical inferences because the statistical testing was to stop after the primary endpoints on the 9.6 mg dose arm. In addition, as noted in the regulatory history, the Agency disagrees with the adequacy of the PADQ as an observer-reported outcome. For the HQ-CT subset and item 9, the Agency noted that such measures could only be considered exploratory without sufficient analysis of psychometric properties and objective support for the meaningfulness of such measures, and that removal of items would affect content validity. In addition, these two endpoints may be redundant to the primary endpoint of the HQ-CT.

Long-Term Follow-Up Period

The LTFU period lacked a placebo control arm. Although study participants (subjects, caregivers, investigators) were blinded to *which* dose of IN carbetocin subjects received (9.6 mg or 3.2 mg), participants were not blinded to receiving active study drug during the LTFU period. Thus, all LTFU period efficacy analyses were subject to confounding by expectation bias, limiting any conclusions that can be drawn from the data.

During the LTFU period, the Applicant conducted efficacy assessments at Weeks 10, 16, 28, 40, 52, 64, and upon Early Termination if applicable. The Applicant asserts that the LTFU period data demonstrate, over time, that subjects who continue to receive IN carbetocin experience meaningful reductions in hyperphagia, obsessive-compulsive, and anxiety and distress symptoms. However, in addition to potential confounding by expectation bias as noted above, the degradation of sample size because of the data cutoff dates and cumulative study discontinuations limits the ability to draw conclusions from the results. For example, for the 3.2 mg arm, the sample size decreased from 59 subjects at Baseline, to 31 subjects at Week 28, to one subject at Week 64. Also, subjects who withdrew may have done so for lack of efficacy, skewing the results.

The Applicant also conducted a prespecified exploratory placebo-crossover analysis for subjects randomized to receive placebo during the 8-week placebo-controlled period, by comparing their changes from Baseline to Week 8 to changes from Week 8 to Week 16 in the LTFU period. According to the Applicant, analysis of subjects who transitioned from placebo to 3.2 mg in the LTFU period resulted in a treatment difference of -9.3190 (95% CI -14.4990, -4.1390) on the HQ-CT. The Applicant asserts that the same comparison for the 9.6 mg dose was not nominally significant because of a "disproportionate placebo response during the first 8 weeks." Of note, the mean change observed for the placebo-to-3.2 mg group during Weeks 8 to 16 on the HQ-CT (-9.0569 [SE 1.6365]) was significantly larger than the mean change from Baseline to Week 8 observed for subjects randomized to 3.2 mg in the placebo-controlled period (-5.372 [SE 0.957]), which suggests confounding by expectation bias in the LTFU period. In addition, the Applicant's analytic approach does not control for confounding by natural history.

Finally, the Applicant conducted a post hoc delayed-start analysis that compared mean changes from Baseline to Week 16 on the HQ-CT for four treatment groups, including those subjects who received the following study drug in the first 8 weeks to the second 8 weeks, respectively: placebo to 9.6 mg, placebo to 3.2 mg, 9.6 mg for all 16 weeks, and 3.2 mg for all 16 weeks. The Applicant reports that all four groups experienced an approximately 9.5-point improvement in mean HQ-CT scores by Week 16. As previously noted, these results are subject to potential confounding by expectation bias, the natural history of PWS, sample size degradation, and potential skewing of results by dropouts for lack of efficacy.

Statistical Considerations and Robustness of Results

The Applicant supports this NDA using exploratory results from the primary analysis. In the pre-specified statistical analysis, the Applicant proposed to first test both endpoints on the high dose (9.6 mg) arm before looking at the endpoints for the low dose (3.2 mg) arm using a multiple comparison procedure (MCP). The Hochberg MCP proposed by the Applicant spent all of the alpha (0.05) on the statistical tests for the primary endpoints and left no alpha for testing endpoints in the 3.2 mg dose arm unless the hypotheses were rejected for both tests on the 9.6 mg dose arm. Once the entire study alpha of 0.05 is used, any additional findings are exploratory (i.e., the type I error is not known). For a confirmatory study, the FDA and ICH E9 recommend controlling the type I error rate to both confirm results of earlier studies and to ensure that only effective drugs are marketed. When a pre-specified statistical testing plan is not followed after seeing the results of a study, the study's Type I error rate cannot be calculated. Any study relying on a *post hoc* revision of the statistical analysis should only be used to generate hypotheses but not to test them.

In the pre-specified primary, there are important assumptions and modeling issues to assess with post hoc sensitivity analyses:

- Compound symmetry covariance structure in the cLDA model
    - Assessment: Explored cLDA models with an unstructured correlation matrix and using a sandwich estimator for the standard errors and how the *p*-values change. In addition, the Agency recommends that all longitudinal data analyses are conducted with an unstructured covariance matrix and the Kenward-Rogers (KR) adjustment to the degrees of freedom adjustment for finite sample sized.
    - Results: Although the *p*-values for the 9.6 mg dose arm changed with the different covariance structures and standard error estimator, the statistical inference did not change because all *p*-values remained greater than 0.05. Changes in the *p*-values for CY-BOCS in the 3.2 mg are not presented for the same reason. In Table 6, note that both the sandwich estimator and unstructured covariance matrix have larger *p*-values compared to the compound symmetry with KR adjustment. This structure has the same results as compound symmetry without the KR adjustment.

Table 6: Study LV-101-3-01 Sensitivity Analysis of Covariance Structure for HQ-CT in the 3.2 mg Arm (Primary Analysis Set)

Covariance Structure / Standard Error Estimator	Treatment Difference	95% CI	*p*-value
Sandwich Estimator[a]	-3.14	-6.24 to -0.03	0.048
Compound Symmetry with Kenward Rogers degrees of freedom	-3.14	-5.69 to 0.58	0.0162
Unstructured	-3.209	-6.36 to -0.058	0.044
Unstructured with Kenward Rogers degrees of freedom	-3.209	-6.46 to 0.044	0.0531

Source: Statistical Reviewer Analysis; adef.xpt
CI = confidence interval
[a] The sandwich estimator is used to correct the standard error estimates from the model using compound symmetry.

- The impact of baseline imbalance on the cLDA model
    - Assessment: Fitted a post hoc mixed model for repeated measures (MMRM) with baseline as a covariate using both compound symmetry and an unstructured covariance matrix.
    - Results: The conditional MMRM analysis yielded p-values of 0.047 and 0.052 for the compound symmetry and unstructured covariance matrix respectively on the HQ-CT endpoint for the 3.2 mg dose arm. All p-values for the 9.6 mg dose arm increased.

These sensitivity analyses indicate that the results of Study LV-101-3-01 are sensitive to violations of the modeling assumptions. This sensitivity to modeling assumptions may be caused by the small per arm sample size (approximately 40 subjects per arm).

The combined statistical issues with the post hoc hypothesis testing and sensitivity to model assumptions lead to a conclusion that Study LV-101-3-01 does not contribute to providing substantial evidence of effectiveness of IN carbetocin.

4. SAFETY OF INTRANASAL CARBETOCIN

4.1. ADVERSE EVENTS AND INVESTIGATIONS

There were no deaths or other serious adverse events (SAEs) during the placebo-controlled treatment periods of Studies 114 and LV-101-3-01.

In Study 114, one subject in the placebo arm discontinued for adverse events (AEs) of agitation, increased aggression, increased hyperphagia, and broken distal ulna. During the placebo-controlled period in Study LV-101-3-01, two subjects in the IN carbetocin 9.6 mg arm discontinued for AEs, including one each for impulsive behavior and tachycardia. In a third subject in the IN carbetocin 9.6 mg arm, an AE of hypersexuality began during the placebo-controlled period but did not lead to discontinuation until the LTFU period.

In Study 114, no AEs reported by ≥ 5% of subjects at a rate greater than placebo occurred in more than one subject. During the placebo-controlled period in Study LV-101-3-01, the most common AEs (≥ 5% incidence and at least twice the rate of placebo) for the IN carbetocin 3.2 mg arm included (in decreasing order of frequency) headache, flushing, diarrhea, abdominal pain and abdominal pain upper (grouped), fatigue, pyrexia, and nasal discomfort. The most common AEs for the IN carbetocin 9.6 mg arm included flushing and epistaxis. See Table 9 in the Appendix for AEs reported by ≥ 3% of subjects in any treatment arm at a rate greater than placebo during the placebo-controlled period.

As of the 120-day safety update data cutoff of June 18, 2021, 17 SAEs occurred for 16 subjects during the LTFU period and ongoing extension of Study LV-101-3-01. SAEs occurring in more than one subject during the LTFU period and extension included scoliosis surgery (four subjects) and pneumonia (two subjects). Nine AEs led to discontinuation for eight subjects during the LTFU period and extension. AEs leading to discontinuation in more than one subject included emotional disorder (two subjects). AEs leading to discontinuation were primarily psychiatric (i.e., in addition to the two subjects who discontinued because of emotional disorder, one subject each discontinued because of aggression, agitation, behavior disorder, obsessive thoughts, and separation anxiety disorder). Discontinuations because of AEs were more common in subjects receiving the 9.6 mg dose (five subjects) versus the 3.2 mg dose (three subjects). During the LTFU and ongoing extension, AEs occurring in ≥5% of subjects for IN carbetocin 3.2 mg included (in decreasing order of frequency) headache, pyrexia, epistaxis, diarrhea, nasopharyngitis, constipation, and anxiety. For IN carbetocin 9.6 mg, AEs occurring in ≥5% of subjects included nasopharyngitis, epistaxis, headache, pyrexia, and anxiety.

No findings for nasal examinations, vital signs, clinical laboratory assessments, or electrocardiograms suggested a safety signal.

Overall, IN carbetocin appeared to be generally safe and well-tolerated in Studies 114 and LV-101-3-01.

5. DOSE-RESPONSE DISCUSSION

Study 114 was smaller and the treatment duration shorter than typical studies intended to demonstrate efficacy and safety in chronic conditions. Nonetheless, the results of Study 114 suggested that the 9.6 mg dose may impact hyperphagia. The primary objective of LV-101-3-01, the subsequent phase 3 study, was to further evaluate the 9.6 mg dose in a longer study with a larger sample size. However, in Study LV-101-3-01, the 9.6 mg dose was not statistically significant on endpoints assessing hyperphagia (or other symptoms) associated with PWS. At the pre-NDA meeting with the Agency, the Applicant hypothesized that the higher dose may have been associated with more off-target vasopressin effects, leading to psychiatric adverse events that may have counteracted the treatment effect. The Applicant has posited that this phenomenon was not as evident in Study 114 because of the shorter duration and that the impact of adverse events emerged with the longer duration of exposure in Study LV-101-3-01. Notably, the 9.6 mg dose did not appear to demonstrate an effect at the end of Week 2 in Study LV-101-3-01, and so results of the phase 3 study did not confirm the phase 2 study results at an equivalent time point. In addition, although there were numerically more adverse events leading to discontinuation in subjects receiving the 9.6 mg dose in the LTFU period, the overall number of events was small, there was no placebo control in the LTFU period for comparison, and it is not clear that the safety profiles of the 3.2 mg and 9.6 mg doses are materially different. Therefore, the Applicant’s hypothesis is not supported by the data.

From a clinical pharmacology perspective, based on the available content submitted to the NDA at this time, there is no apparent mechanistic rationale to explain why the efficacy is greater for the 3.2 mg TID compared to 9.6 mg TID dose level in the Phase 3 trial LV-101-3-01.

From a statistical standpoint, no formal statistical assessment of the secondary endpoints could be conducted in Study LV-101-3-01 given the lack of statistical significance on the primary endpoints. However, exploratory analyses of the secondary endpoints—including an evaluation of the effect of the lower dose (3.2 mg)—were conducted to better understand carbetocin’s potential effects. Despite small descriptive *p*-values for some secondary endpoints, there is uncertainty about the effectiveness of carbetocin, particularly given the aforementioned inconsistency in the phase 2 and phase 3 study results, the lack of additional adequate and well-controlled substantiating studies, and lack of robustness in the statistical findings. Conclusions regarding the efficacy of carbetocin in the LTFU period were limited by the absence of a placebo control, the potential for expectation bias, and degradation of the sample size over time.

As discussed during the development program, further clinical studies designed to evaluate the effects of the 3.2 mg dose would be informative, particularly because treatments for hyperphagia in PWS would potentially be administered indefinitely starting in early childhood. The severity of hyperphagia may vary over time and may be affected by the types of behavioral interventions and environmental controls that are in place, complicating efforts of patients, caregivers, and clinicians to discern the effectiveness once treatment has been initiated.

6. REFERENCES

1. Driscoll DJ, Miller JL, Schwartz S, and Cassidy SB, Prader-Willi Syndrome, 1998 Oct 6 [Updated 2017 Dec 14], In: Adam MP, Ardinger HH, Pagon RA, et al., editors, GeneReviews [Internet], Seattle (WA): University of Washington, Seattle; 1993-2021, Available from: https://www.ncbi.nlm.nih.gov/books/NBK1330/

2. Cassidy SB, Schwartz S, Miller JL, and Driscoll DJ, 2012, Prader-Willi Syndrome, Genet Med, 14:10-26.

3. Miller JL, Lynn CH, Driscoll DC, Goldstone AP, Gold J, Kimonis V, Dykens E, Butler MG, Shuster JJ, and Driscoll DJ, 2011, Nutritional Phases in Prader-Willi Syndrome, Am J Med Genet A, 155A(5):1040-1049.

4. Wharton RH, Wang T, Graeme-Cook F, Briggs S, and Cole RE, 1997, Acute Idiopathic Gastric Dilation With Gastric Necrosis in Individuals With Prader-Willi Syndrome, Am J Med Genet, 73(4):437-431.

5. Stevenson DA, Heineman J, Angulo M, Butler MG, Loker J, Rupe N, Kendell P, Clericuzio CL, and Scheimann AO, 2007, Deaths Due to Choking in Prader-Willi Syndrome, Am J Med Genet A, 143A(5):484-487.

6. Butler MG, Manzardo AM, Heinemann J, Loker C, and Loker J, 2017, Causes of Death in Prader-Willi Syndrome: Prader-Willi Syndrome Association (USA) 40-Year Mortality Survery, Genet Med, 19(6):635-642.

7. Manzardo AM, Loker J, Heinemann J, Loker C, and Butler MG, 2018, Survival Trends From the Prader-Willi Association (USA) 40-Year Mortality Survey, Genet Med, 20(1):24-30.

8. Scheimann AO, Butler MG, Miller JL, Lee PD, Stevenson DA, Heinemann J, and Driscoll DJ, 2012, Letter to the Editor: Long-Term Experience With Duodenal Switch in Adolescents. Obes Surg, 22(3):517-518.

9. Holland AJ, Aman LCS, and Whittington JE, 2019, Defining Mental and Behavioral Disorder in Genetically Determined Neurodevelopmental Syndrome With Particular Reference to Prader-Willi Syndrome, Genes, 10(12):1025-1039.

10. Dykens E and Shah B, 2003, Psychiatric Disorders in Prader-Willi Syndrome: Epidemiology and Management, CNS Drugs, 17(3):167–78.

7. APPENDICES

Table 7: Schedule of Assessments, Study 000114

Period:	Screening	Treatment					F/U	
		Baseline			1-Week Assessment		End of Treatment Visit / Early termination visit	F/U phone call
Visit:	1	2	3				4	
Days:	-7 to 0	1	2	3-7	8 (+/-1)	9-14	15	19 (+/-3)
IMP Dosing:	NA	On site	On site	Off site			On site	NA
Assessments								
Written informed consent and assent	X							
Inclusion/exclusion criteria	X							
Prior/concomitant medication	X	X	X		X		X	
Physical examination	X						X	
Nasal examination[a]	X	X	X				X	
Medical history	X							
Demographics	X							
Urine drug and alcohol screen	X						X	
Urine pregnancy test	X						X	
Height and Weight	X						X	
Vital signs[b]	X	X	X				X	
Electrocardiogram[c]	X	X	X				X	
Blood samples for clinical chemistry, hematology, hemostasis, and urinalysis	X		X				X	
Blood samples for pharmacokinetics[d]		X	X					
Nutritional phase assessment	X							
Randomization[e]		X[e]						
HPWSQ-R and HPWSQ-R-C	X	X			X		X	
CGI-S		X						
CGI-I					X		X	
CY-BOCS	X				X		X	
Food Domain of Reiss Profile	X				X		X	
IMP administration training		X						
Dispense IMP		X	X	X	X	X		
Return IMP							X	
IMP dosing		X	X	X	X	X		
Adverse events	X	X	X		X		X	X

(continued next page)

CGI-I=Clinical Global Impression-Improvement after treatment; CGI-S=Clinical Global Impression-Severity; CY-BOCS= Children's Yale-Brown Obsessive-Compulsive scale; F/U=follow-up; HPWSQ-R=Hyperphagia in Prader-Willi Syndrome Questionnaire-Responsiveness; HPWSQ-R-C=Hyperphagia in Prader-Willi Syndrome Questionnaire-Clinician ; IMP=investigational medicinal product

[a] Nasal examinations to be done at Visit 1 (anytime), within 1 hour prior to each dose at Visit 2 and Visit 3, and at Visit 4 (anytime).

[b] Includes pulse, respiration rate, blood pressure and body temperature. At Visit 2 and 3, vital signs will be monitored prior to and after IMP dosing at 30-minute intervals for 2 hours after each dose.

[c] ECG to be done within 1 hour after morning dose at Visit 2 and 3.

[d] Three blood samples to be collected at Visit 2 and Visit 3, according to the schedule specified in section 7.2.6 Clinical Laboratory Tests/Pharmacokinetic Assessments. All PK labs are to be done prior to subsequent dose.

[e] Randomization will be done within 24 hours prior to the Visit 2, morning dose.

Source: Study 114 Clinical Study Report, Table 5-1, p. 22 (including footnotes)

Table 8: Schedule of Assessments, 8-Week Placebo-Controlled Period, Study LV-101-3-01

Procedure	Screening (14 to 28 days Prior to Baseline)	Baseline Site Visit	Week 2 Site Visit (±2 days)	Week 8 Site Visit (±2 days)
Informed consent/assent	X			
Inclusion and exclusion criteria	X	X		
Demography	X			
Medical history	X	X		
Physical examination[a]	X	X	X	X
Nasal assessment	X	X	X	X
Vital signs (including weight)[a]	X	X	X	X
Height[a]	X			
Pregnancy test[b]	X	X	X	X
Laboratory assessments (chemistry, hematology, coagulation, urinalysis)[a]	X	X	X	X
12-lead ECG[a]	X	X		X
Randomization (eligible subjects)		X[c]		X[d]
Training on IP administration		X		
Study drug dosing		X	ongoing	ongoing[e]
AE review		X	X	X
SAE review	X	X	X	X
Concomitant medication review	X	X	X	X
Caregiver training on clinical outcomes assessments	X	X		
Structured interview, including clinical outcomes assessments by caregiver and clinician[f]	X	X	X	X
Plasma collection for PK			X[g]	
Archive blood/plasma samples[h]		X	X	X
Questionnaires				
PWS Nutritional Phase Assessment	X	X		
SRS-2[i]		X		X
Food Safe Zone	X			
HQ-CT	X	X	X	X
H-CGI and H-CareGI	X[j]	X[j]	X[j]	X[j]
CY-BOCS	X	X	X	X

(Continued next page)

Procedure	Screening (14 to 28 days Prior to Baseline)	Baseline Site Visit	Week 2 Site Visit (±2 days)	Week 8 Site Visit (±2 days)
OC-CGI and OC-CareGI	X[j]	X[j]	X[j]	X[j]
PADQ	X	X	X	X
A-CGI and A-CareGI	X[j]	X[j]	X[j]	X[j]
CGI	X[j]	X[j]	X[j]	X[j]

A = anxiety; AE = adverse event; CareGI = Caregiver Global Impression; CGI = Clinical Global Impression; CY-BOCS = Children's Yale-Brown Obsessive-Compulsive Scale; ECG = electrocardiogram; H = hyperphagia; HQ-CT = Hyperphagia Questionnaire for Clinical Trials; IN = intranasal; IP = investigational product; OC = obsessive-compulsive; PADQ = PWS Anxiety and Distress Behaviors Questionnaire; PK = pharmacokinetics; PWS = Prader-Willi syndrome; SAE = serious AE; SRS-2 = Social Responsiveness Scale, Second Edition

[a] In the case of a remote visit, physical examinations, vital signs, laboratory draws, and ECGs could be deferred.

[b] Pregnancy testing was conducted only for females of childbearing potential, as judged by the Investigator. In the case of a remote visit, urine pregnancy tests for females of child-bearing potential were to be performed at the Investigator's discretion, as applicable and possible.

[c] Randomization occurred 1 day prior to Baseline (in order to be able to thaw frozen LV-101).

[d] At the time of initial randomization, subjects who were randomized to placebo in the placebo-controlled period were further randomized (1:1) to 3.2 mg/dose LV-101 three times per day before meals or 9.6 mg/dose LV-101 three times per day before meals starting at the Week 8 visit.

[e] In the case of a remote visit, the Week 8 study dosing should have occurred only after all remote visit procedures had been completed. The Investigator used his/her discretion if the Week 8 dosing was to be observed by the Investigator or Study Coordinator.

[f] Refer to Questionnaires section of table. Appropriate training and procedures were instituted (including structured interviews) to ensure familiarity with the instruments, that the instruments were consistently administered (reducing measurement error), and to enable honest feedback.

[g] A blood sample was obtained 25 to 35 minutes after observed dosing of study drug. In the case of a remote visit, plasma collection for PK was not conducted.

[h] Archived samples were collected for potential future immunogenicity testing (presence and characterization of anti-drug antibodies), genetic testing (including genome sequencing), and biochemical analyses of hormones or other analytes of relevance to PWS. In the case of a remote visit, no archive blood/plasma samples were collected.

[i] In the case of a remote visit, the SRS-2 questionnaire was completed last. If there was not enough time to complete this questionnaire during the remote visit, the site may have chosen to have the caregiver complete the questionnaire separately after the remote visit (preferably the same day) and return the completed form via mail/scan/email to the site.

[j] Screening and baseline assessments included Severity only, while post-baseline assessments included Severity and Change.

Source: Study LV-101-3-01 Clinical Study Report, Table 4, p. 48 (including footnotes)

Table 9: Schedule of Assessments, Long-Term Follow-Up Period (LTFU), Study LV-101-3-01

Procedure	Week 10 (±2 Days)	Week 16 (±1 Week)	Week 28 (±1 Week)	Week 40 (±1 Week)	Week 52 (±1 Week)	Week 64 (±1 Week)	Early Term (If Required)
Physical examination [a]	X	X	X	X	X	X	X
Nasal assessment	X	X	X	X	X	X	X
Vital signs (including weight) [a]	X	X	X	X	X	X	X
Height [a]						X	X
Pregnancy test [b]	X	X	X	X	X	X	X
Laboratory assessments (chemistry, hematology, coagulation, urinalysis) [a]	X	X	X	X	X	X	X
12-lead ECG [a]		X				X	X
Study drug dosing	ongoing	ongoing	ongoing	ongoing	ongoing	X	
Extension period determination						X	
AE review	X	X	X	X	X	X	X
SAE review	X	X	X	X	X	X	X
Concomitant medication review (including all supplements)	X	X	X	X	X	X	X
Structured interview, including clinical outcomes assessments by caregiver and clinician [c]	X	X	X	X	X	X	X
Plasma collection for PK	X[d]						
Archive blood/plasma samples [e]	X	X	X	X	X	X	X
Questionnaires							
PWS Nutritional Phase Assessment						X	X
Food Safe Zone						X	X
HQ-CT	X	X	X	X	X	X	X
H-CGI and H-CareGI	X[f]	X[f]	X[f]	X[f]	X[f]	X[f]	X[f]
CY-BOCS	X	X	X	X	X	X	X
OC-CGI and OC-CareGI	X[f]	X[f]	X[f]	X[f]	X[f]	X[f]	X[f]
PADQ	X	X	X	X	X	X	X
A-CGI and A-CareGI	X[f]	X[f]	X[f]	X[f]	X[f]	X[f]	X[f]
CGI	X[f]	X[f]	X[f]	X[f]	X[f]	X[f]	X[f]
SRS-2[g]		X				X	X

A = anxiety; AE = adverse event; CareGI = Caregiver Global Impression; CGI = Clinical Global Impression; CY-BOCS = Children's Yale-Brown Obsessive-Compulsive Scale; ECG = electrocardiogram; H = hyperphagia; HQ-CT = Hyperphagia Questionnaire for Clinical Trials; OC = obsessive-compulsive; PADQ = PWS Anxiety and Distress Behaviors Questionnaire; PK = pharmacokinetics; PWS = Prader-Willi syndrome; SAE = serious adverse event; SRS-2 = Social Responsiveness Scale, Second Edition

[a] In the case of a remote visit, physical examinations, vital signs, laboratory draws, and ECGs could be deferred.

[b] Pregnancy testing was conducted only for females of childbearing potential, as judged by the Investigator. In the case of a remote visit, urine pregnancy tests for females of child-bearing potential were to be performed at the Investigator's discretion, as applicable and possible.

(Continued next page)

[c] Refer to Questionnaires section of table. Appropriate training and procedures will be instituted (including structured interviews) to ensure familiarity with the instruments, administer instruments consistently (reducing measurement error), and enable honest feedback.

[d] Obtained at 5 to 10 minutes, 25 to 35 minutes, 1.5 hours ± 15 minutes, and 2.5 hours ± 15 minutes after observed dosing. In the case of a remote visit, plasma collection for PK was not conducted.

[e] Archive samples were collected for potential future immunogenicity testing (presence and characterization of anti-drug antibodies), genetic testing (including genome sequencing) and biochemical analyses of hormones or other analytes of relevance to PWS. In the case of a remote visit, no archive blood/plasma samples were collected.

[f] Screening and baseline assessments included Severity only, while post-baseline assessments included Severity and Change.

[g] In the case of a remote visit, the SRS-2 questionnaire was completed last. If there was not enough time to complete this questionnaire during the remote visit, the site may have chosen to have the caregiver complete the questionnaire separately after the remote visit (preferably the same day) and return the completed form via mail/scan/email to the site.

Source: Study LV-101-3-01 Clinical Study Report, Table 5, p. 50 (including footnotes)

Hyperphagia in Prader-Willi Syndrome Questionnaire-Responsiveness (HPWSQ-R)

This survey will help keep track of how things have been with your child in the past week.

For each one of the survey items, please mark an ☒ in the one box that best describes your answer.

1. These questions are about how things have been with your child during the past week? For each question, please give the one answer that comes closest to the way to describe the situation-

	Not at all	A little bit	Moderate	Quite a bit	Extremely
a How upset did your child generally become when denied a desired food?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
b How persistent was your child in asking or looking for food after being told "no" or "no more"?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
c How distressed did your child become when stopped from talking about food or engaging in food-related behaviors?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
d How "clever" or "fast" was your child in obtaining food?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
e To what extent did food-related thoughts, talk, or behavior interfere with your child's normal daily routines, self-care, school, or work?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

(Continued next page)

2. How much of the time during the past week did your child …

	None of the time	A little of the time	Some of the time	Most of the time	All of the time
a Try to bargain or manipulate to get more food	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
b Forage through the trash for food	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
c Get up at night to food seek?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
d Talk about food or engaged in food-related behaviors, outside of normal meal times	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
e Try to steal food	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

3. During the past week, once your child had food on his/her mind, how easy was it for you or others to re-direct your child away from food to other things?

Extremely easy	Very easy	Somewhat hard	Very hard	Extremely hard
☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

Source: Study 000114 Protocol, Appendix II

Hyperphagia Questionnaire for Clinical Trials (HQ-CT)

Instructions:
The following items refer to the person in your care and assessment of his/her food-related behavior during the past 2 weeks.

(1) During the past 2 weeks, how upset did the person generally become when denied a desired food?

☐ Not at all upset

☐ A little upset

☐ Moderately upset

☐ Very upset

☐ Extremely upset

(2) During the past 2 weeks, how often did the person try to bargain or manipulate to get more food at meals?

☐ Never

☐ Up to 2 times a week

☐ 3 to 6 times a week

☐ Every day

☐ Several times a day

(3) During the past 2 weeks, how often did the person forage through trash for food?

☐ Never

☐ 1 time

☐ 2 times

☐ 3 times

☐ 4 or more times

(4) During the past 2 weeks, how often did the person get up at night to food seek?

☐ Never

☐ 1 time

☐ 2 times

☐ 3 times

☐ 4 or more times

(Continued next page)

(5) During the past 2 weeks, how persistent was the person in asking or looking for food after being told "no" or "no more"?

- ☐ Not at all persistent
- ☐ A little persistent
- ☐ Moderately persistent
- ☐ Very persistent
- ☐ Extremely persistent

(6) During the past 2 weeks, outside of normal meal times, how much time did the person generally spend asking or talking about food?

- ☐ Less than 5 minutes a day
- ☐ 5 to 15 minutes a day
- ☐ 15 to 30 minutes a day
- ☐ 30 minutes to 1 hour a day
- ☐ More than 1 hour a day

(7) During the past 2 weeks, how often did the person try to sneak or steal food (that you are aware of)?

- ☐ Never
- ☐ 1 time
- ☐ 2 times
- ☐ 3 times
- ☐ 4 or more times

(8) During the past 2 weeks, when others tried to stop the person from asking about food, how distressed did he or she generally appear?

- ☐ Not at all distressed
- ☐ A little distressed
- ☐ Moderately distressed
- ☐ Very distressed
- ☐ Extremely distressed

(9) During the past 2 weeks, how often did food-related behavior interfere with the person's normal daily activities, such as self-care, recreation, school, or work?

- ☐ Never
- ☐ Up to 2 times a week
- ☐ 3 to 6 times a week
- ☐ Every day
- ☐ Several times a day

Source: Study LV-101-3-01 Protocol, Questionnaires 12.1

Children's Yale-Brown Obsessive-Compulsive Scale

CY-BOCS Symptom Checklist
Children's Yale-Brown Obsessive Compulsive Scale
Administering the CY-BOCS Symptom Checklist and CY-BOCS Severity Ratings

1. Establish the diagnosis of obsessive compulsive disorder.
2. Using the CY-BOCS Symptom Checklist (below), ascertain current and past symptoms.
3. Next, administer the 10 item severity ratings (other form) to assess the severity of the OCD during the last week.
4. Re-administer the CY-BOCS Severity Rating Scale to monitor progress.

Patient ____ Date ________________

(Continued next page)

CY-BOCS Obsessions Checklist

Check all symptoms that apply (Items marked "*" may or may not be OCD Phenomena)

Current	Past	Contamination Obsessions	Current	Past	Sexual Obsessions
□	□	Concern with dirt, germs, certain illnesses (e.g., AIDS)	□	□	Forbidden or perverse sexual thoughts, images, impulses
□	□	Concerns or disgust with bodily waste or secretions (e.g. urine, feces, saliva)	□	□	Content involves homosexuality
□	□	Excessive concern with environmental contaminants (e.g., asbestos, radiation, toxic waste)	□	□	Sexual behavior towards others (aggressive)
□	□	Excessive concern with household items (e.g., cleaners, solvents)	□	□	Other (describe) ____________
□	□	Excessive concern about animals / insects			**Hoarding / Saving Obsessions**
□	□	Excessively bothered by sticky substances or residues	□	□	Fear of losing things
□	□	Concerned will get ill because of contaminant	□	□	Other (describe) ____________
□	□	Concerned will get others ill by spreading contaminant (aggressive)	□	□	**Magical Thoughts / Superstitious Obsessions**
□	□	No concern with consequences of contamination other than how it might feel *	□	□	Lucky / unlucky numbers, colors, words
□	□	Other (describe) ____________	□	□	Other (describe) ____________
		Aggressive Obsessions			**Somatic Obsessions**
□	□	Fear might harm self	□	□	Excessive concern with illness or disease *
□	□	Fear might harm others	□	□	Excessive concern with body part or aspect of appearance (e.g. dismorphophobia)
□	□	Fear harm will come to self	□	□	Other (describe) ____________
□	□	Fear harm will come to others (maybe because of something child did or did not do)			**Religious Obsessions**
□	□	Violent or horrific images	□	□	Excessive concern or fear of offending religious objects
□	□	Fear of blurting out obscenities or insults	□	□	Excessive concern with right / wrong morality
□	□	Fear of doing something embarrassing *	□	□	Other (describe) ____________
□	□	Fear will act on unwanted impulses (e.g., to stab a family member)			**Miscellaneous Obsessions**
□	□	Fear will steal things	□	□	The need to know or remember
□	□	Fear will be responsible for something else terrible happening (e.g. ,fire, burglary, flood)	□	□	Fear of saying certain things

(Continued next page)

□	□	Other (describe) ____________	□	□	Fear of not saying just the right thing
			□	□	Intrusive (non – violent) images
			□	□	Intrusive sounds, words, music, or numbers
			□	□	Other (describe) ____________

Target Symptoms List for Obsessions

OBSESSIONS (describe, listing by order of severity, with #1 being most severe, #2 second most severe, etc):

1. ______________________________
2. ______________________________
3. ______________________________
4. ______________________________

(Continued next page)

CY-BOCS Symptom Checklist
Children's Yale-Brown Obsessive Compulsive Scale

CY-BOCS Compulsions Checklist
Check all symptoms that apply (Items marked "*" may or may not be OCD phenomena)

Current	Past	Washing / Cleaning Compulsions	Current	Past	Hoarding / Saving Compulsions
□	□	Excessive or ritualized hand washing			Distinguish from hobbies and concern for objects of monetary or sentimental value
□	□	Excessive or ritualized showering, bathing, tooth brushing, grooming, toilet routine	□	□	Difficulty throwing things away, saving bits of paper, string, etc.
□	□	Excessive cleaning of items, such as personal clothes or important objects	□	□	Other (describe) ___________
□	□	Other measures to prevent or remove contact with contaminants			**Excessive Games / Superstitious Behaviors**
□	□	Other (describe) ___________	□	□	Distinguish from age appropriate magical games (e.g. array of behavior, such as sleeping over certain spots on a floor, touching an object / self a certain number of times to prevent something bad from happening)
		Checking Compulsions	□	□	Other (describe) ___________
□	□	Checking locks, toys, school books / items, etc.	□	□	**Rituals Involving Other Persons**
□	□	Checking associated with getting washed, dressed, or undressed	□	□	The need to involve another person (usually a parent) in ritual (e.g. asking a parent to repeatedly answer the same question, making mother perform certain mealtime rituals involving specific utensils)
□	□	Checking that did not / will not harm others	□	□	Other (describe) ___________
□	□	Checking that did not / will not harm self	□	□	**Miscellaneous Compulsions**
□	□	Checking that nothing terrible did / will happen	□	□	Mental rituals other than checking / counting
□	□	Checking that did not make mistake	□	□	Need to tell, ask, or confess
□	□	Checking tied to somatic obsessions	□	□	Measures, not checking, to prevent:
□	□	Other (describe) ___________	□	□	harm to self
		Repeating Rituals	□	□	harm to others
□	□	Rereading, erasing, or rewriting	□	□	terrible consequences
□	□	Need to repeat activities (e.g. in / out of doorway, up / down from chair)	□	□	Ritualized eating behaviors*

(Continued next page)

□	□	Other (describe) ________	□	□	Excessive list making*
		Counting Compulsions	□	□	Need to touch, tap, rub*
□	□	Objects, certain numbers, words, etc.	□	□	Need to do things (e.g. touch or arrange until it feels just right)
□	□	Other (describe) ________	□	□	Rituals involving blinking or staring
		Ordering / Arranging	□	□	Trichotillomania (hair pulling)
□	□	Need for symmetry / evening up (e.g. lining items up in a certain way or arranging personal items in specific patterns)	□	□	Other self-damaging or self-mutilating behaviors
			□	□	Other (describe) ________

Target Symptoms List for Compulsions

COMPULSIONS (describe, listing by order of severity, with #1 being the most severe, #2 second most severe, etc.)

1. ________________________________
2. ________________________________
3. ________________________________
4. ________________________________

(Continued next page)

CY-BOCS Severity Ratings

Children's Yale-Brown Obsessive Compulsive Scale

Administering the CY-BOCS Symptom Checklist and CY-BOCS Severity Ratings

1. Establish the diagnosis of obsessive compulsive disorder.
2. Using the CY-BOCS Symptom Checklist (other form), ascertain current and past symptoms.
3. Next, administer the 10-item severity ratings (below) to assess the severity of the OCD during the last week.
4. Readminister the CY-BOCS Severity Rating Scale to monitor progress.

Patient ______________________________

Date 1st Report ______________________________ Date This Report ______________________________

Obsession Rating Scale (circle appropriate score)

Note: Scores should reflect the composite effect of all the patient's obsessive compulsive symptoms.
Rate the average occurrence of each item during the prior week up to and including the time of interview.

QUESTIONS ON OBSESSIONS (ITEMS 1-5) *"I AM NOW GOING TO ASK YOU QUESTIONS ABOUT THE THOUGHTS YOU CANNOT STOP THINKING ABOUT."*
(Review for the informant(s) the Target Symptoms and refer to them while asking questions 1-5).

1. Time Occupied by Obsessive Thoughts
[Be sure to exclude ruminations and preoccupations which, unlike obsessions, are ego-syntonic and rational (but exaggerated)]

	None	Mild	Moderate	Severe	Extreme
		less than 1 hr/day or occasional intrusion	1 to 3 hrs/day or frequent intrusion	greater than 3 and up to 8 hrs/day or very frequent intrusion	greater than 8 hrs/day or near constant intrusion
Score	0	1	2	3	4

2. Interference Due to Obsessive Thoughts
- How much do these thoughts get in the way of school or doing things with friends?
- Is there anything that you don't do because of them? (If currently not in school, determine how much performance would be affected if patient were in school)

	None	Mild	Moderate	Severe	Extreme
		slight interference with social or school activities, but overall performance not impaired	definite interference with social or school performance, but still manageable	causes substantial impairment in social or school performance	incapacitating
Score	0	1	2	3	4

3. Distress Associated with Obsessive Thoughts

	None	Mild	Moderate	Severe	Extreme
		infrequent, and not too disturbing	frequent, and disturbing, but still manageable	very frequent, and very disturbing	near constant, and disabling distress/frustration
Score	0	1	2	3	4

4. Resistance Against Obsessions
- How hard do you try to stop the thoughts or ignore them? (Only rate effort made to resist, not success or failure in actually controlling the obsessions. If the obsessions are minimal, the patient may not feel the need to resist them. In such cases, a rating of "0" should be given.)

	None	Mild	Moderate	Severe	Extreme
	makes an effort to always resist, or symptoms so minimal doesn't need to actively resist	tries to resist most of the time	makes some effort to resist	yields to all obsessions without attempting to control them, but does so with some reluctance	completely and willingly yields to all obsessions
Score	0	1	2	3	4

5. Degree of Control Over Obsessive Thoughts

	Complete Control	Much Control	Moderate Control	Little Control	No Control
		usually able to stop or divert obsessions with some effort and concentration	sometimes able to stop or divert obsessions	rarely successful in stopping obsessions, can only divert attention with difficulty	experienced as completely involuntary, rarely able to even momentarily divert thinking
Score	0	1	2	3	4

Obsession subtotal (add items 1-5) ______________________________

(Continued next page)

QUESTIONS ON COMPULSIONS (ITEMS 6-10) *"I AM NOW GOING TO ASK YOU QUESTIONS ABOUT THE HABITS YOU CAN'T STOP"*
(Review for the informant(s) the Target Symptoms and refer to them while asking questions 6-10)

6. Time Spent Performing Compulsive Behaviors

	None	Mild less than 1 hr/day	Moderate 1 to 3 hrs/day	Severe greater than 3 & up to 8 hrs/day	Extreme greater than 8 hrs/day
Score	0	1	2	3	4

7. Interference Due to Compulsive Behaviors
- How much do these habits get in the way of school or doing things with friends?
- Is there anything you don't do because of them? (If currently not in school, determine how much performance would be affected if patient were in school.)

	None	Mild slight interference with social or school activities, but overall performance not impaired	Moderate definite interference with social or school performance, but still manageable	Severe causes substantial impairment in social or school performance	Extreme incapacitating
Score	0	1	2	3	4

8. Distress Associated with Compulsive Behavior
- How would you feel if prevented from carrying out your habits? How upset would you become?

	None	Mild only slightly anxious if compulsions prevented	Moderate anxiety would mount but remain manageable if compulsions prevented	Severe prominent and very disturbing increase in anxiety if compulsions interrupted	Extreme incapacitating anxiety from any intervention aimed at modifying activity
Score	0	1	2	3	4

9. Resistance Against Compulsions
- How much do you try to fight the habits? (Only rate effort made to resist, not success or failure in actually controlling the compulsions.)

	None makes an effort to always resist, or symptoms so minimal doesn't need to actively resist	Mild tries to resist most of the time	Moderate makes some effort to resist	Severe yields to all obsessions without attempting to control them, but does so with some reluctance	Extreme completely and willingly yields to all obsessions
Score	0	1	2	3	4

10. Degree of Control Over Compulsive Thoughts
- How strong is the feeling that you have to carry out the habit(s)?
- When you try to fight them, what happens?

	Complete Control	Much Control experiences pressure to perform the behavior, but usually able to exercise voluntary control over it	Moderate Control moderate control, strong pressure to perform behavior, can control it only with difficulty	Little Control little control, very strong drive to perform behavior, must be carried to completion, can only delay with difficulty	No Control no control, drive to perform behavior experienced as completely involuntary and overpowering, rarely able to delay activity (even momentarily)
Score	0	1	2	3	4

Compulsion subtotal (add items 6-10) ____________

CY-BOCS total (add items 1-10) ____________

Total CY-BOCS score: range of severity for patients who have both obsessions and compulsions

0-7	**Subclinical**	**24-31**	**Severe**
8-15	**Mild**	**32-40**	**Extreme**
16-23	**Moderate**		

Source: Study 000114 Protocol, Appendix IV

Table 10: Adverse Events Reported by ≥3% of Subjects in Any Treatment Group at a Rate Greater Than Placebo During the Placebo-Controlled Period, Study LV-101-3-01

MedDRA System Organ Class Preferred Term	IN carbetocin 3.2 mg (N=43) n (%)	IN carbetocin 9.6 mg (N=44) n (%)	Placebo (N=43) n (%)
At least one adverse event	26 (60.5%)	29 (65.9%)	24 (55.8%)
Gastrointestinal disorders	9 (20.9%)	3 (6.8%)	6 (14.0%)
Diarrhea	4 (9.3%)	2 (4.5%)	1 (2.3%)
Abdominal pain, abdominal pain upper	3 (7.0%)	1 (2.3%)	1 (2.3%)
Constipation	2 (4.7%)	1 (2.3%)	1 (2.3%)
General disorders and administrative site conditions	8 (18.6%)	4 (9.1%)	1 (2.3%)
Pyrexia	3 (7.0%)	--	--
Fatigue	3 (7.0%)	1 (2.3%)	--
Feeling hot	1 (2.3%)	2 (4.5%)	--
Infections and infestations	6 (14.0%)	4 (9.1%)	10 (23.3%)
Upper respiratory tract infection	3 (7.0%)	2 (4.5%)	2 (4.7%)
Sinusitis	2 (4.7%)	1 (2.3%)	1 (2.3%)
Injury, poisoning and procedural complications	4 (9.3%)	2 (4.5%)	3 (7.0%)
Skin abrasion	2 (4.7%)	1 (2.3%)	1 (2.3%)
Metabolism and nutrition disorders	2 (4.7%)	2 (4.5%)	--
Hyperphagia	2 (4.7%)	2 (4.5%)	--
Nervous system disorders	11 (25.6%)	7 (15.9%)	5 (11.6%)
Headache	7 (16.3%)	4 (9.1%)	3 (7.0%)
Dizziness	2 (4.7%)	2 (4.5%)	1 (2.3%)
Psychiatric disorders	6 (14.0%)	8 (18.2%)	4 (9.3%)
Anxiety	--	2 (4.5%)	--
Dermatillomania	2 (4.7%)	1 (2.3%)	--
Respiratory, thoracic and mediastinal disorders	8 (18.6%)	9 (20.5%)	2 (4.7%)
Epistaxis	1 (2.3%)	6 (13.6%)	1 (2.3%)
Nasal discomfort	3 (7.0%)	2 (4.5%)	1 (2.3%)
Oropharyngeal pain	2 (4.7%)	--	--
Vascular disorders	6 (14.0%)	9 (20.5%)	--
Flushing	6 (14.0%)	9 (20.5%)	--

Source: Clinical Reviewer-created from 120-Day Safety Update Report, Table 4, p. 10, and ADAE dataset
MedDRA = Medical Dictionary for Regulatory Activities (Version 23.1)
Note: Subjects may have had more than one adverse event in a System Organ Class or Preferred Term, but were only counted once for each.

8. ATTACHMENT

FDA Draft Guidance: Demonstrating Substantial Evidence of Effectiveness for Human Drug and Biological Products Guidance for Industry (2019)

Demonstrating Substantial Evidence of Effectiveness for Human Drug and Biological Products

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document, contact (CDER) Ei Thu Lwin, Office of New Drug Policy, 301-796-0728 or (CBER) Office of Communication, Outreach and Development, 800-835-4709 or 240-402-8010, ocod@fda.hhs.gov.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Biologics Evaluation and Research (CBER)
Center for Drug Evaluation and Research (CDER)

December 2019
Clinical/Medical

Demonstrating Substantial Evidence of Effectiveness for Human Drug and Biological Products

Guidance for Industry

Additional copies are available from:
Office of Communications, Division of Drug Information
Center for Drug Evaluation and Research
Food and Drug Administration
10001 New Hampshire Ave., Hillandale Bldg., 4th Floor
Silver Spring, MD 20993-0002
Phone: 855-543-3784 or 301-796-3400; Fax: 301-431-6353
Email: druginfo@fda.hhs.gov
https://www.fda.gov/drugs/guidance-compliance-regulatory-information/guidances-drugs

and/or

Office of Communication, Outreach and Development
Center for Biologics Evaluation and Research
Food and Drug Administration
10903 New Hampshire Ave., Bldg. 71, Room 3128
Silver Spring, MD 20993-0002
Phone: 800-835-4709 or 240-402-8010
Email: ocod@fda.hhs.gov
https://www.fda.gov/vaccines-blood-biologics/guidance-compliance-regulatory-information-biologics

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Biologics Evaluation and Research (CBER)
Center for Drug Evaluation and Research (CDER)

December 2019
Clinical/Medical

TABLE OF CONTENTS

Demonstrating Substantial Evidence of Effectiveness for Human Drug and Biological Products Guidance for Industry

> This draft guidance, when finalized, will represent the current thinking of the Food and Drug Administration (FDA or Agency) on this topic. It does not establish any rights for any person and is not binding on FDA or the public. You can use an alternative approach if it satisfies the requirements of the applicable statutes and regulations. To discuss an alternative approach, contact the FDA staff responsible for this guidance as listed on the title page.

I. INTRODUCTION

This document is intended to provide guidance to applicants planning to file new drug applications (NDAs), biologics license applications (BLAs), or applications for supplemental indications on the evidence to be provided to demonstrate effectiveness. This guidance complements and expands on the 1998 guidance entitled Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products (the 1998 guidance)[1].

The 1998 guidance was issued in response to the Food and Drug Administration Modernization Act of 1997 (FDAMA) (Pub. L. 105–115), which stated that the substantial evidence requirement for effectiveness, which had generally been interpreted as calling for two adequate and well-controlled trials, could also be met by a single trial[2] plus confirmatory evidence. The 1998 guidance, therefore, provided many examples of the types of evidence that could be considered confirmatory evidence, with a specific focus on adequate and well-controlled trials of the test agent in related populations or indications, as well as a number of illustrations of a single adequate and well-controlled trial supported by convincing evidence of the drug's mechanism of action in treating a disease or condition.

FDAMA thus introduced a specific new area of flexibility in the evidence needed to support effectiveness, but there are many other characteristics of the evidence supporting effectiveness that can vary (notably, trial designs, trial endpoints, statistical methodology), and evidence that varies in such ways potentially can provide substantial evidence of effectiveness but because of these characteristics may provide greater or lesser certainty. These characteristics also deserve consideration and were not discussed in the 1998 guidance. FDA's consideration of these various designs, endpoints, and analyses which can differ in the strength of evidence they

---

[1] FDA updates guidances periodically. To make sure you have the most recent version of a guidance, check the FDA webpage. The guidances mentioned in this document are available on the guidance web page at https://www.fda.gov/drugs/guidance-compliance-regulatory-information/guidances-drugs, and https://www.fda.gov/vaccines-blood-biologics/guidance-compliance-regulatory-information-biologics.

[2] In this guidance, the terms "trial" and "clinical trial" have the same meaning as the term "clinical investigation" as the latter is defined in FDA regulations (see, e.g., 21 CFR 312.3(b)).

provide, reflects the Agency's longstanding flexibility when considering the types of data and evidence that can meet the substantial evidence requirement.

Although FDA's evidentiary standard for effectiveness has not changed since 1998, the evolution of drug development and science has led to changes in the types of drug development programs submitted to the Agency. Specifically, there are more programs studying serious diseases lacking effective treatment, more programs in rare diseases, and more programs for therapies targeted at disease subsets. There is a need for more Agency guidance on the flexibility in the amount and type of evidence needed to meet the substantial evidence standard in these circumstances. The approaches discussed in this guidance can yield evidence that meets the statutory standard for substantial evidence and reflect the evolving landscape of drug development.

The "substantial evidence" of effectiveness standard in the statute (discussed in Section II) refers to both the quality and the quantity of the evidence. It clearly provides that all clinical investigations supporting effectiveness should be of appropriate design and of high quality (i.e., adequate and well-controlled; discussed in Section III). Sponsors often seek advice on what trial design will be considered acceptable in various development programs. This guidance discusses, in part, what clinical trial designs are considered adequate and well-controlled, and under what circumstances it may be appropriate to use a given design (discussed in Section III.A).

The clinical endpoints studied are a critical aspect of evidence quality (discussed in Section III.B). The Agency accepts clinical endpoints that reflect patient benefits (i.e., how patients feel, function, or survive) or validated surrogate endpoints[3] (i.e., those that have been shown to predict a specific clinical benefit) as the basis for traditional approval. In contrast to traditional approval, accelerated approval can be based on a demonstrated effect on a surrogate endpoint that is reasonably likely to predict a clinical benefit but where there are not sufficient data to show that it is a validated surrogate endpoint. Effects on intermediate clinical endpoints[4] can also be a basis for accelerated approval. For drugs granted accelerated approval, FDA requires post-approval trials to verify the predicted clinical benefit.

This guidance also discusses the quantity of evidence needed in a given development program – i.e., two adequate and well-controlled trials, one adequate and well-controlled trial plus confirmatory evidence, or reliance on a previous finding of effectiveness of an approved drug when scientifically justified and legally permissible (i.e., no new effectiveness or pharmacodynamic data would be needed) (discussed in the 1998 guidance and Section IV.A, IV.B, and IV.C, respectively). It also expands upon the discussions included in the 1998 guidance on the types of mechanistic and pharmacologic evidence and non-clinical evidence that can constitute confirmatory evidence.

---

[3] For more information on validated surrogate endpoints, see the BEST (Biomarkers, EndpointS, and other Tools) Resource available at: https://www ncbi nlm.nih.gov/books/NBK453484/.

[4] An intermediate clinical endpoint is "a clinical endpoint that can be measured earlier than irreversible morbidity or mortality, that is reasonably likely to predict an effect on irreversible morbidity or mortality or other clinical benefit." Section 506(c)(1)(A) of the Federal Food, Drug, and Cosmetic Act (FD&C Act).

Although randomized superiority trials with a placebo- or active-control design generally provide the strongest evidence of effectiveness, this guidance discusses the circumstances under which trials not using a placebo control, superiority design, or randomization may be acceptable (discussed in Section V.A and V.B). In addition, this guidance also discusses situations in which human efficacy trials are not ethical or feasible, and the animal rule may be applied (discussed in Section V.C).

The finding of substantial evidence of effectiveness is necessary but not sufficient for FDA approval. The approval decision also requires a determination that the drug is safe for the intended use. As all drugs have adverse effects, evaluating whether a drug is "safe" involves weighing whether the benefits of the drug outweigh its risks under the conditions of use defined in labeling. Uncertainties regarding benefits and risks are considered when making an approval determination; a drug with greater risks may require a greater magnitude and certainty of benefit to support approval. This benefit-risk analysis, as well as other determinations necessary for approval, is outside the scope of this guidance.

In general, FDA's guidance documents do not establish legally enforceable responsibilities. Instead, guidances describe the Agency's current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word *should* in Agency guidances means that something is suggested or recommended, but not required.

II. STANDARD OF EFFECTIVENESS FOR DRUGS AND BIOLOGICS

A. Statutory standard

In 1962, Congress required for the first time that drugs be shown to be effective as well as safe. A drug's effectiveness must be established by "substantial evidence," which is defined as:

> "evidence consisting of adequate and well-controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved, on the basis of which it could fairly and responsibly be concluded by such experts that the drug will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in the labeling or proposed labeling thereof."[5]

Under section 351 of the Public Health Service Act (PHS Act) (42 U.S.C.§ 262) licenses for biologics have been issued only upon a showing that the products are "safe, pure, and potent." Potency has long been interpreted to include effectiveness (21 CFR 600.3(s)). FDA has also generally considered "substantial evidence" of effectiveness to be necessary to support licensure

[5] The FD&C Act section 505(d) (21 U.S.C. § 355(d)).

of a biological product under section 351 of the PHS Act.[6]

FDA has interpreted the law as generally requiring at least two adequate and well-controlled clinical investigations,[7] each convincing on its own, to establish effectiveness (discussed in Section IV.A.1). Under specific circumstances, however, FDA has considered a large multicenter trial that has certain characteristics to satisfy the legal requirement for substantial evidence of effectiveness (discussed in Section II.C.3 of the 1998 guidance and Section IV.A.2). FDA may also rely on a previous finding of effectiveness of an approved drug when scientifically justified and legally permissible; in this case there is no need for additional adequate and well-controlled clinical efficacy trials (discussed in Section IV.C).

In addition to reliance on a single large multicenter trial or previous finding of effectiveness of an approved drug, there are other circumstances where substantial evidence of effectiveness can be provided outside of the setting of two adequate and well-controlled clinical investigations. Congress specifically provided for these in section 115(a) of FDAMA, which amended the statutory provision on substantial evidence of effectiveness, 21 U.S.C. § 355(d), to add the following:

> "If [FDA] determines, based on relevant science, that data from one adequate and well-controlled clinical investigation and confirmatory evidence (obtained prior to or after such investigation) are sufficient to establish effectiveness, [FDA] may consider such data and evidence to constitute substantial evidence."

This modification explicitly recognized the potential for FDA to find that one adequate and well-controlled clinical investigation with confirmatory evidence, including supportive data outside of a controlled trial, is sufficient to establish effectiveness (discussed in Section IV.B).

B. Scientific basis for the statutory standard

To establish a drug's effectiveness, it is essential to distinguish the effect of the drug "from other influences, such as spontaneous change in the course of the disease, placebo effect, or biased observation." [8] This is the basis for the statutory requirement that approval be based on adequate and well-controlled investigations, as well as the basis for FDA's regulations describing the characteristics of such investigations (i.e., design elements that are generally intended to minimize bias and permit a valid comparison with a control to provide a quantitative assessment of drug effect).

---

[6] In 1972, FDA initiated a review of the safety and effectiveness of all previously licensed biologics. The Agency stated then that proof of effectiveness would, with limited exceptions, consist of controlled clinical investigations as defined in the provision for "adequate and well-controlled studies" for new drugs (21 CFR 314.126) (see former 21 CFR 601.25(d)(2) (2015) (revoked as no longer necessary, 81 FR 7445 (Feb. 12, 2016))). We note that, in section 123(f)) of FDAMA, Congress also directed the agency to take measures to "minimize differences in the review and approval" of products required to have approved BLAs under section 351 of the PHS Act and products required to have approved NDAs under section 505(b)(1) of the FD&C Act.

[7] See FDA regulation regarding adequate and well-controlled studies at 21 CFR 314.126.

[8] 21 CFR 314.126(a).

A second adequate and well-controlled investigation or confirmatory evidence provides substantiation of experimental results, which is a widely accepted scientific principle. This approach is intended to minimize the possibility that other influences such as bias and chance findings could result in a false conclusion that a drug is effective when in fact it is not (false positive).

III. THE QUALITY OF CLINICAL EVIDENCE TO ESTABLISH EFFECTIVENESS

The quality of clinical evidence to establish effectiveness and the resulting level of certainty about the demonstration of substantial evidence is impacted by the selection of trial design and trial endpoint(s) as well as statistical considerations, as discussed below.

A. Trial designs

Adequate and well-controlled clinical investigations provide the primary basis for determining whether there is substantial evidence to support the claims of effectiveness.[9] FDA regulation at 21 CFR 314.126(b) describes characteristics of an adequate and well-controlled clinical investigation, including choice of control, method of patient assignment to treatment (e.g., randomization), adequate measures to minimize bias (e.g., blinding), well-defined and reliable assessment of individuals' response (i.e., efficacy endpoint), and adequate analysis of the clinical investigation's results to assess the effects of the drug (i.e., statistical methods). Although randomized double-blinded, concurrently controlled superiority trials are usually regarded as the most rigorous design, as discussed further below, five types of controls are described in section 314.126:[10] placebo concurrent control, dose-comparison concurrent control, no treatment concurrent control, active treatment concurrent control, and historical control (a type of external control).[11] Of note, when the first version of the rule was published in 1970, historical controls and active treatment controls were included.[12] Thus, from its earliest description of adequate and well-controlled trials, FDA included trial designs (as discussed below) that may be more difficult to interpret, which reflected FDA's recognition that different trial designs (including choice of control) may be appropriate in different disease settings.

Establishing superiority to a concurrent control group (whether an active agent, including a lower dose of the test drug, or placebo) generally provides strong evidence of effectiveness, because a superiority design does not depend on assumptions regarding the effectiveness of the control.

[9] The FD&C Act section 505(d) (21 U.S.C. § 355(d)); 21 CFR 314.126(a).
[10] See 50 FR 7452, 7487 (February 22, 1985).
[11] The regulation uses the term "historical control," which is a subset of "external control." FDA also accepts other types of external controls. An externally controlled trial compares a group of subjects receiving the test treatment with a group of patients external to the trial, rather than to an internal control group consisting of patients from the same trial population assigned to a different treatment. The external control can be a group of patients, treated or untreated, at an earlier time (historical control) or a group, treated or untreated, during the same time period but in another setting. An important subset of externally controlled trials are "baseline controlled trials," where there is not a specific external control group but assurance, based on experience, that no change could occur (e.g., tumors are known not to shrink spontaneously or patients not given general anesthetic remain awake). See International Conference on Harmonisation E10 guidance on Choice of Control Group and Related Issues in Clinical Trials (ICH E10). This guidance uses the term "external control," except when referring to section 314.126.
[12] See 35 FR 7250, 7251-7252 (May 8, 1970).

However, each of the trial designs has distinct considerations; for example, the lack of blinding when using a no treatment control could introduce bias, which may attenuate confidence in the trial's results. The dose-comparison design may support the effectiveness of the highest dose when a positive dose response is seen, but could leave uncertainty about whether lower tested doses were effective.

Although demonstrating that a new drug is superior to an active control provides strong evidence of effectiveness, a common goal of active controlled trials is to show non-inferiority (NI), i.e., that the new drug is not less effective than the active control by a specified amount, that amount being no larger than the effect the active control was expected (the effect is not measured) to have had in the NI trial based on the drug's past performance in trials. Showing such non-inferiority allows a conclusion that the new drug is effective.[13] In general, with regard to establishing effectiveness, NI designs are credible and appropriate only in situations in which the active control has shown a consistent effect (generally compared with placebo) in prior superiority trials conducted in a patient population similar to the population in the clinical investigation being planned. Unless a placebo group (or other treatment group where the intent is to demonstrate superiority of the test drug) is also included, these NI trials depend on the assumption, not confirmed in the trial, that the active control had its anticipated effect (which is the basis for the NI margin) in the trial. As a result, the strength of evidence that may result from an NI trial can vary considerably depending on the specific disease setting and the choice of active control. An NI trial that meets its objective (with respect to the pre-specified statistical testing plan) could mean either that both drugs were effective or, if neither control nor drug has its expected effect, that neither was effective in the trial. Because interpretation of NI trials depends on assumptions not confirmed in the trial, this design is usually chosen when it would be unethical or infeasible to conduct one of the superiority designs discussed above (e.g., when withholding available therapy would not be clinically acceptable and the new drug is being studied as an alternative, rather than as an adjunct, to available therapy).

Externally controlled trials differ in several important ways from the other trial designs identified in 21 CFR 314.126. Most notably, random assignment is not a feature of external control designs. As a result, there may be differences in patient characteristics or concomitant treatments in the trial population compared to the external control population that lead to differences in outcomes that are unrelated to the investigational treatment. In addition, the lack of blinding could introduce bias. For these reasons, external control designs are usually reserved for specific circumstances, such as trials of diseases with high and predictable mortality or progressive morbidity (e.g., certain malignancies or certain rare diseases) and trials in which the effect of the drug is self-evident (e.g., general anesthetics).

Despite the limitations of externally controlled trials compared with concurrently controlled trials, strong support for effectiveness can emerge from externally controlled trials, especially when (1) the natural history of a disease is well defined, (2) the external control population is very similar to that of the treatment group, (3) concomitant treatments that affect the primary endpoint are not substantially different between the external control population and the trial population, and (4) the results provide compelling evidence of a change in the established progression of disease. Such results could include partial or complete response in a disease

[13] FDA guidance on Non-Inferiority Clinical Trials to Establish Effectiveness.

where spontaneous regression is not observed, or stabilization or improvement in function in a disease where progressive functional decline is well documented to occur over the duration of the treatment period in the trial. Another example of where there is strong evidence of drug effectiveness is reversal of clinical signs and symptoms following a toxic exposure or overdose after administration of a drug antidote. In all such circumstances, a detailed understanding of the full range of possible clinical outcomes, with a well-documented natural history of the disease in the absence of treatment, is essential to interpreting trial results and, therefore, drawing a conclusion about the effectiveness of the drug.

It is important to recognize that trial design alone does not determine whether evidence from the trial is sufficient to establish substantial evidence of effectiveness. For example, compelling results may overcome challenges associated with less rigorous trial designs, such as those with an external control. As discussed above, a small externally controlled trial with an outcome markedly superior to the well-established natural history of a disease may provide a compelling case for drug effectiveness. Similarly, a successful active-controlled NI trial of a new antimicrobial drug or of a new anticoagulant to prevent stroke in patients with atrial fibrillation can provide strong evidence of effectiveness when it is well-established that the effect of the control antimicrobial or anticoagulant drug is large.

Poor execution can render a trial of any design to be not adequate or not well-controlled and, therefore, unable to provide substantial evidence of effectiveness. Examples of this include (1) a randomized, double-blind, placebo-controlled trial where there is extensive drop-out of trial patients (with the potential for informative censoring), and (2) a randomized, double-blind, placebo-controlled trial in which unblinding is common due to an effect of the test drug, and where a modest treatment effect is found on a primary endpoint that is subject to bias when drug assignment is known (e.g., a physician global impression). In these cases, the trials might not be considered adequate and well-controlled.

B. Trial endpoints

One of the characteristics of an adequate and well-controlled clinical investigation is that "the methods of assessment of subjects' response are well-defined and reliable."[14] Such a method of assessment can be a clinical endpoint[15] or, where appropriate, a surrogate endpoint.[16]

---

[14] 21 CFR 314.126(b)(6).

[15] An endpoint is a precisely defined variable intended to reflect an outcome of interest as a measure of drug effect that is prespecified (i.e., chosen before the data are analyzed) and statistically analyzed to address a particular research question. A definition of "clinical endpoint" is provided in FDA guidance on Expedited Programs for Serious Conditions – Drugs and Biologics (FDA guidance on expedited programs). A clinical endpoint can be used to support traditional approval.

[16] A definition of "surrogate endpoint" is provided in FDA guidance on expedited programs. A surrogate endpoint that has been shown to predict a specific clinical benefit can be used to support traditional approval. A surrogate endpoint that is reasonably likely to predict clinical benefit can be used to support accelerated approval. Accelerated approval can also be based on an effect on a clinical endpoint that can be measured earlier than irreversible morbidity or mortality (IMM) and that is reasonably likely to predict an effect on IMM or other clinical benefit. See FDA web page on Table of Surrogate Endpoints That Were the Basis of Drug Approval or Licensure, available at https://www fda.gov/Drugs/DevelopmentApprovalProcess/DevelopmentResources/ucm613636.htm.

Although the statutory standard for effectiveness does not refer to particular endpoints or state a preference for clinical endpoints over surrogate endpoints, it is well established that the effect shown in the adequate and well-controlled clinical investigations, must be, in FDA's judgment, clinically meaningful.[17]

Many disease specific guidances have been issued by the Agency that can assist sponsors in identifying an appropriate trial endpoint. In addition, discussion with appropriate review divisions early in clinical development can assist sponsors in identifying appropriate trial endpoints for a particular development program.

C. Statistical considerations

The strength of evidence in each trial contributing to meeting the substantial evidence standard should be assessed by appropriate statistical methods. The uncertainty about the findings from each trial should be sufficiently small and the findings should be unlikely to result from chance alone, as demonstrated by a statistically significant result or a high posterior probability of effectiveness.[18] Statistical approaches should be specified in advance, to limit erroneous conclusions resulting from multiplicity.

IV. THE QUANTITY OF CLINICAL EVIDENCE TO ESTABLISH EFFECTIVENESS

A. Meeting the substantial evidence standard based on two adequate and well-controlled clinical investigations

1. Two adequate and well-controlled clinical investigations

In many situations FDA requires two adequate and well-controlled trials to establish effectiveness. This reflects the need for substantiation of experimental results, which has often been referred to as the need for replication of the finding. Replication may not be the best term, however, as it may imply that precise repetition of the same experiment in other patients by other investigators is the only means to substantiate a conclusion. Although two positive identically designed and conducted trials can provide substantial evidence of effectiveness, precise replication of a trial is only one of a number of possible means of obtaining substantiation of a clinical finding and, at times, can provide less persuasive evidence of benefit, as it could leave the conclusions of both trials vulnerable to any systematic biases inherent to the particular study design.

Two positive trials with differences in design and conduct may be more persuasive, as unrecognized design flaws or biases in study conduct will be less likely to impact the outcomes of both trials. The consistency of results across two trials also greatly reduces the possibility that a biased, chance, site-specific, or fraudulent result will lead to an erroneous conclusion that a

---

[17] See preamble to FDA final rule on accelerated approval (57 FR 58942, 58944 (December 11, 1992)).

[18] In a Bayesian framework the strength of evidence is assessed by the probability that the drug is effective given the data rather than by statistical significance.

drug is effective. Such trials also may be more informative: for example, two positive trials using the same endpoint but with distinct study populations within the same proposed indication (e.g., one trial studying a new glucose-lowering drug in patients with type 2 diabetes receiving only diet and exercise therapy, and a second trial in patients with type 2 diabetes already on two or three oral antihyperglycemic agents) may provide evidence that is more generalizable to the population that will take the drug than two identical trials in a narrower population. Similarly, two trials in the same disease using different but related clinical endpoints could support effectiveness and provide broader information about the drug's effect (e.g., one trial showing symptom improvement and a second trial showing improved survival in a more severely ill population).

2. One adequate and well-controlled large multicenter trial that can provide substantial evidence of effectiveness

In general, substantiation of a drug's effectiveness obtained with two trials, especially with complementary design, as discussed above, will provide more convincing evidence of effectiveness than would a single trial. In some circumstances, however, there may not be a meaningful difference between the strength of evidence provided by a single large multicenter adequate and well-controlled trial and that provided by two smaller adequate and well-controlled trials. In such cases, the large multicenter trial can be considered, both scientifically and legally, to be, in effect, multiple trials and can be relied on to provide substantial evidence of effectiveness. Large multicenter trials can include a broad range of subjects and investigation sites and have procedures in place to ensure trial quality (e.g., investigation site selection, monitoring, and auditing). They generally are less vulnerable to certain biases such as selection or measurement bias, are often more generalizable to the intended population, and can often be evaluated for internal consistency across subgroups, centers, and multiple endpoints.

Reliance on a single large multicenter trial to establish effectiveness should generally be limited to situations in which the trial has demonstrated a clinically meaningful and statistically very persuasive effect on mortality, severe or irreversible morbidity, or prevention of a disease with potentially serious outcome, and with other characteristics described below, and confirmation of the result in a second trial would be impracticable or unethical. For example, conducting a second trial after a strongly positive trial had demonstrated a decrease in post-infarction mortality, or prevention of pertussis would generally present significant ethical concerns. Repetition of positive trials showing only symptomatic benefit would generally not present the same ethical concerns.

In addition to the expectation that the single trial is large and multicenter, there should be no single trial site that is the main contributor to the observed effect, either by virtue of having a much bigger effect or many more patients than other sites; these characteristics help address concerns about bias and chance findings associated with a single trial. As noted above it would also be expected that the effect size on the primary endpoint and the statistical analysis results are both persuasive.

Other characteristics, discussed below, also support the persuasiveness of a single trial in supporting the conclusion that there is substantial evidence of effectiveness. Finding consistent, and clinically meaningful effects on distinct prospectively specified endpoints (e.g., an effect on both myocardial infarction and stroke for a drug being studied for cardiovascular benefit) can provide further evidence that the results are not due to chance. Moreover, an effect on a meaningful, objective endpoint, such as certain imaging endpoints, may complement a more subjective endpoint, such as a clinician- or patient-reported outcome. In these cases, the internal consistency across endpoints not only reduces the possibility of a chance finding but also may further support the clinical utility of the results.

Frequently, large multicenter trials have relatively broad entry criteria and the trial populations may be diverse with regard to important covariates such as concomitant or prior therapy, disease stage, age, gender or race. Analysis of the results of such trials for consistency across important patient subgroups can address concerns about generalizability of findings to various populations in a manner that may not be possible with smaller trials or trials with more narrow entry criteria.

Furthermore, there may be other characteristics of a large multicenter trial that increase confidence in its results. For example, the multicenter trial may sometimes be appropriately analyzed as "multiple trials" within a single trial. An example is a 4-arm ("2×2 factorial") trial (placebo, drug A, drug B, and drug A + drug B) in which the effectiveness of drug A could be supported by two controlled comparisons if the combination of drug A + drug B is superior to drug B alone *and* drug A is superior to placebo.

Although a large multicenter trial with robust results can be persuasive, even a robust result can arise from bias. For example, although two consistent findings within a single trial usually provide reassurance that a positive treatment effect is not due to chance, they do not protect against bias in trial conduct, biased analyses, or fraud. Thus, close scrutiny of trial conduct, including, for example, completeness of follow-up, methods of analysis, imputation of missing data, evaluation of trial endpoints, is critical to evaluating such trials. Findings from other trials that are not consistent with the findings of the single positive trial would need to be considered collectively, and could weaken the overall strength of evidence.

B. Meeting the substantial evidence standard based on one adequate and well-controlled clinical investigation plus confirmatory evidence

Under certain circumstances and consistent with FDAMA, FDA can conclude that one adequate and well-controlled clinical investigation plus confirmatory evidence is sufficient to establish effectiveness. FDA will consider a number of factors when determining whether reliance on a single adequate and well-controlled clinical investigation plus confirmatory evidence is appropriate. These factors may include the persuasiveness of the single trial; the robustness of the confirmatory evidence; the seriousness of the disease,[19] particularly where there is an unmet medical need; the size of the patient population; and whether it is ethical and practicable to conduct more than one adequate and well-controlled clinical investigation. Sponsors intending to establish substantial evidence of effectiveness using one adequate and well-controlled clinical

[19] While seriousness of the disease is one of the factors that FDA considers, reliance on a single trial plus confirmatory evidence to establish effectiveness is not limited only to drugs for "serious diseases," as the term is defined in 21 CFR 312.300(b)(1).

investigation plus confirmatory evidence should consult FDA in advance to discuss the appropriateness of such an approach for their development program.

Confirmatory evidence could include, for example, adequate and well-controlled clinical investigations in a related disease area, certain types of real world evidence[20] such as extensive data on outcomes that provide further support for the lack of effect seen in the control group in the randomized trial, compelling mechanistic evidence in the setting of well-understood disease pathophysiology (e.g., pharmacodynamic data or compelling data from nonclinical testing), or well-documented natural history of the disease.

Below are examples of when a single adequate and well-controlled clinical investigation, together with confirmatory evidence, can establish effectiveness. The strength of the single trial will affect the extent of confirmatory evidence required – for example, a trial showing compelling efficacy results (but not rising to the level that would be provided by a large multicenter trial, as discussed in Section IV.A.2) may require less confirmatory evidence.

1. One adequate and well-controlled clinical investigation on a new indication for an approved drug, supported by existing adequate and well-controlled clinical investigation(s) that demonstrated the effectiveness of the drug for its other, closely related approved indication(s)

To establish effectiveness for a new indication of a product already approved by FDA – where the new indication is closely related to the other approved indication(s) – substantial evidence of effectiveness can be based on one adequate and well-controlled clinical investigation, generally a randomized concurrently controlled trial, of the new indication, supported by the confirmatory evidence provided by the existing adequate and well-controlled clinical investigation(s) that established effectiveness of the product for the related indication(s). See Section II.C.2 of the 1998 guidance for more details.

2. One adequate and well-controlled clinical investigation supported by data that provide strong mechanistic support

A single adequate and well-controlled clinical investigation, generally a randomized concurrently controlled trial, together with earlier phase clinical results and/or testing that provide compelling mechanistic evidence in the setting of well-understood disease pathophysiology, may be sufficient to provide substantial evidence of effectiveness of a new drug or a new indication. The mechanistic evidence would generally be obtained from clinical testing using a relevant and well understood pharmacodynamic endpoint not accepted by itself as an endpoint to establish evidence of effectiveness. It also could be collected from other sources, such as animal studies (e.g., those using an established, relevant animal model to study the effect of the drug on a pharmacodynamic marker of known relevance to humans), or a combination of

[20] Real world evidence is the clinical evidence regarding the usage, and potential benefits or risks, of a medical product derived from analysis of real world data. Real world data are data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources. See FDA real world evidence web page, available at https://www fda.gov/science-research/science-and-research-special-topics/real-world-evidence.

the two.[21] An example is enzyme replacement therapy, where a single adequate and well-controlled clinical investigation that demonstrates the therapy's efficacy is supported by evidence that the condition is caused by the enzyme deficiency and by earlier results that show the therapy increases enzyme activity to biologically active levels at the appropriate site and/or reduces disease-specific substrates. Another example could be a trial of a drug which is a mineral or vitamin replacement that showed restoration of accepted normal concentrations, in concert with a prior large body of information showing the clinical consequences of deficiency states.

3. One adequate and well-controlled clinical investigation with compelling results, supported by additional data from the natural history of the disease

In certain circumstances, FDA accepts one adequate and well-controlled clinical investigation that has generated compelling results as the basis to demonstrate effectiveness, when the single trial is supported by additional data from the natural history of the disease that reinforce the very persuasive finding. For example, a single trial showing marked improvement in survival compared to a control group, either external to the trial or concurrent, could be supported by data from separate sources (e.g., a natural history study, case report forms, or registries) that demonstrate a very limited median survival time or other clinically highly important outcome without treatment. In this case, the natural history data would represent confirmatory evidence.

4. One adequate and well-controlled clinical investigation of the new drug, supported by scientific knowledge about the effectiveness of other drugs in the same pharmacological class

In certain circumstances, FDA accepts one adequate and well-controlled clinical investigation as the basis to demonstrate effectiveness, when the single trial is supported by confirmatory evidence of effectiveness from adequate and well-controlled trials of other drugs in the same pharmacological class.[22] For example, the approval of two angiotensin II receptor blockers, losartan and irbesartan, for the treatment of diabetic nephropathy in patients with type 2 diabetes, hypertension, and abnormal kidney function, was based on effectiveness data from a single trial of each drug, supported by similarly favorable results from a single trial of the other drug. In this

---

[21] FDA supports the principles of the "3Rs," to reduce, refine, and replace animal use in testing when feasible. FDA encourages sponsors to consult with us if they wish to use a non-animal testing method they believe is suitable, adequate, validated, and feasible. FDA will consider if such an alternative method could be assessed for equivalency to an animal test method.

[22] Reliance on data concerning a different drug raises legal issues that will need to be considered in each case. If the applicant owns the data concerning the other drug, or has a right to refer to those data, such as a license, then the legal concerns are satisfied. In the example of losartan and irbesartan cited in the text, the two applicants each agreed to permit the other to rely on their data. If there is not such permission, for an NDA, the question will be raised whether the reliance makes the application a 505(b)(2) application. If so, that may require compliance with patent certification requirements applicable to such applications and may mean that the submission or approval of the application will be affected by statutory exclusivity provisions. For a BLA, in certain circumstances reliance on data not owned by the applicant, that is not in the public domain, and for which the applicant does not have a right of reference would raise additional legal considerations.

case, the two single trials supplied the needed confirmatory evidence for each other, as neither drug would have been approved for this indication based on the single trial alone.[23]

Whether this scenario applies to a particular development program depends on a number of factors, including but not limited to: (1) the strength of the evidence for effectiveness from the single trial; and (2) the relevance of the additional data derived from other drugs in the same class, including the similarity between the new drug and other drugs in the same class, particularly the pharmacologic activity or specificity of mechanism of action.[24]

C. Meeting the substantial evidence standard for a new population or a different dose, regimen, or dosage form, based on reliance of FDA's previous finding of effectiveness of an approved drug when scientifically justified and legally permissible

When scientifically justified and legally permissible, FDA can rely on its previous finding of effectiveness of an approved drug to conclude that the drug "will have the effect it purports or is represented to have,"[25] thus not requiring additional adequate and well-controlled clinical efficacy trials. Ordinarily, this will be because other types of evidence provide a way to apply the known effectiveness to a new population or a different dose, regimen, or dosage form. For example, the effectiveness of a drug for pediatric use can sometimes be based on FDA's previous finding of effectiveness of the drug in adults, together with scientific evidence that justifies such reliance.[26] In this case, the scientific evidence may include, for example, evidence supporting a conclusion of similar disease course and pathophysiologic basis in adult and pediatric populations, and similar pharmacologic activity of the drug in adults and children (e.g., similar concentration-response relationships), as well as similar blood levels of the drug in adults and children. The effectiveness of new dosage forms or dosing regimens may be demonstrated by the effectiveness trial(s) on the original dosage form or regimen, together with evidence that both the dosage forms or regimens have similar pharmacokinetic (PK) profiles. In this case no new effectiveness or pharmacodynamic data would be needed, but sufficient safety data would still be needed. See Section II.C.1. of the 1998 guidance for more details.

---

[23] See Secondary Review Memo on losartan, May 3, 2002, available at https://www.accessdata.fda.gov/drugsatfda_docs/nda/2002/20386-S028_COZAAR_Medr1.pdf; see also the FDA-approved labels for both products.

[24] A product development program under this scenario may result in a small safety database. Sponsors should consult FDA guidance on Premarketing Risk Assessment, which notes that the appropriate size of a safety database depends on a number of factors specific to the product; two of them are particularly relevant to this scenario, i.e., the product's novelty (i.e., whether it represents a new treatment or is similar to available treatment) and the availability of alternative therapies and the relative safety of those alternatives as compared to the new product. For more details, see FDA guidance on Premarketing Risk Assessment.

[25] See the statutory definition of "substantial evidence" in section 505(d) of the FD&C Act.

[26] Section 505B(a)(2)(B)(i) of the FD&C Act (21 U.S.C. § 355c(a)(2)(B)(i)).

V. EXAMPLES OF CLINICAL CIRCUMSTANCES WHERE ADDITIONAL FLEXIBILITY MAY BE WARRANTED

The statutory standard of "substantial evidence" contains both a statement of what kind of evidence must exist ("adequate and well-controlled investigations") and also an element of expert judgment. Thus the standard requires that the investigations be such that "it could fairly and responsibly be concluded by [qualified] experts that the drug will have the effect it purports or is represented to have,"[27] and permits approval on the basis of one trial and confirmatory evidence only "If [FDA] determines, based on relevant science, that data . . . are sufficient to establish effectiveness." For example, while FDA regulations outline five different types of studies that might be considered adequate and well-controlled,[28] it has always been recognized that some designs (e.g., placebo concurrent control) provide more certainty than others (e.g., external controls). FDA experts may "fairly and responsibly" rely on study designs that produce less certainty in some circumstances when a better design is not feasible or ethical. This may be the case for life-threatening and severely debilitating diseases with an unmet medical need, for certain rare diseases, or potentially even for a more common disease where the availability of existing treatments makes certain design choices infeasible or unethical. FDA would not, however, find it responsible to rely on such design choices in other situations in which, for example, the drug will be used for a less serious disease and greater certainty about benefits and risks is needed, or in cases where designs providing more certainty are possible. In all cases, FDA must reach the conclusion that there is substantial evidence of effectiveness to approve a drug; however, the degree of certainty supporting such a conclusion may differ, depending on clinical circumstances (e.g., severity and rarity of the disease and unmet medical need).

This reflects the longstanding awareness that, in certain settings, a somewhat greater risk (compared to placebo-controlled or other randomized superiority trials) of false positive conclusions – and therefore less certainty about effectiveness – may be acceptable, when balanced against the risk of rejecting or delaying the marketing of an effective therapy, as described below for an unmet medical need. The data supporting effectiveness could, despite the greater risk of error, support a conclusion that there is substantial evidence of effectiveness. Therefore, when selecting a trial design, a sponsor should consider the specific clinical circumstance, including the severity of the disease, unmet medical need (e.g., whether there is available therapy), the rarity of the disease, and whether it is feasible and ethical to conduct a randomized concurrently controlled superiority trial.

A. When the disease is life-threatening or severely debilitating with an unmet medical need

As defined in 21 CFR 312, subpart E (21 CFR 312.81), the term "life-threatening" means diseases or conditions where the likelihood of death is high unless the course of the disease is interrupted, and diseases or conditions with potentially fatal outcomes, where the endpoint of clinical trial analysis is survival; the term "severely debilitating" means diseases or conditions

[27] The law is clear that it is the FDA which "must determine, after giving full consideration to all of the evidence that has been submitted, including expert opinions, if the studies meet the regulatory criteria and show effectiveness." Warner-Lambert Co. v. Heckler, 787 F.2d 147, 154 (3rd Cir. 1986).
[28] 21 CFR 314.126(b)(2).

that cause major irreversible morbidity. An unmet medical need is a condition whose treatment or diagnosis is not addressed adequately by available therapy.[29]

Subpart E regulations promulgated in 1988[30] call for FDA to exercise its broad scientific judgment in applying the evidentiary approval standards to drugs for life-threatening and severely debilitating diseases, especially where there is no satisfactory alternative therapy. In addition, the accelerated approval regulations built upon this recognition by acknowledging that reliance on a surrogate endpoint "almost always introduces some uncertainty into the risk/benefit assessment, because clinical benefit is not measured directly and the quantitative relation of the effect on the surrogate to the clinical effect is rarely known."[31] Together these regulations recognize the importance of facilitating the development of, and access to, safe and effective treatment options for life-threatening and severely debilitating diseases with unmet medical needs. This approach has been reinforced by FDA's interactions with patients and their caregivers who describe their willingness to accept less certainty about effectiveness in return for earlier access to much needed medicines. For example, for a life-threatening disease without any available treatment, FDA might accept the results of adequate and well-controlled investigations with less rigorous designs, such as a historically controlled study. Below are considerations for drugs developed for life-threatening and severely debilitating diseases.

1. Trial design

While a randomized placebo-controlled trial can provide more definitive evidence of a small treatment effect than any other kind of trial of the same size, there are instances when this design and other concurrently controlled superiority designs may not be feasible or ethical. In such settings, other trial designs, such as non-inferiority trials or externally controlled trials can be acceptable if they provide substantial evidence of effectiveness (see discussion of noninferiority design and external control in Section III.A).

2. Trial endpoints

As discussed in Section III.B, endpoint selection is an important consideration in clinical trial design. The most straightforward and readily interpreted endpoints are those that directly measure clinical benefit or are validated surrogate endpoints shown to predict clinical benefit. Surrogate endpoints that are reasonably likely to predict clinical benefit can be relied on to establish effectiveness under the accelerated approval pathway. Effects on intermediate clinical endpoints can also be a basis for accelerated approval. Surrogate and intermediate clinical endpoints often can be assessed sooner than an endpoint that directly measures the clinical benefit or irreversible morbidity or mortality. Note that for accelerated approval the evidentiary standard still applies – that is, there must be substantial evidence that the drug has a meaningful effect on the surrogate or intermediate clinical endpoint.

---

[29] FDA guidance on expedited programs.

[30] 21 CFR 312.80, subpart E; 21 CFR 314.105(c).

[31] The preamble to the final rule on accelerated approval also notes, when responding to a comment, that "[a]lthough studies using surrogate endpoints may provide less assurance of clinical benefit than studies using clinical endpoints, FDA believes compliance with all of the elements of the accelerated approval program will not result in the marketing of large numbers of clinically ineffective drugs." 57 FR 58942, 58944 (December 11, 1992).

3. Number of trials

Although two adequate and well-controlled clinical investigations remain the standard approach to generating substantial evidence of effectiveness in many disease settings, there are scenarios where the conduct of a second trial is not ethical or feasible.

For example, as discussed in section IV.A.2, when a large multicenter trial has demonstrated a clinically meaningful and statistically very persuasive effect on mortality, irreversible morbidity, or prevention of a disease with potentially serious outcome, a second trial would be impracticable or unethical. In this case the single large multicenter trial would be considered sufficient to establish effectiveness.

4. Statistical considerations

A typical criterion for concluding that a trial is positive (showed an effect) is a p value of < 0.05 (two sided). A lower p value, for example, would often be expected for reliance on a single trial. For a serious disease with no available therapy or a rare disease where sample size might be limited, as discussed further below, a somewhat higher p value – if prespecified and appropriately justified – might be acceptable.

B. When the disease is rare

By statutory definition, a rare disease – including a genetically defined subset of a disease – affects fewer than 200,000 people in the U.S.;[32] but many rare diseases affect far fewer patients. A large number of rare diseases are pediatric diseases or have childhood onset. In addition, many rare disorders are life-threatening or severely debilitating diseases with no approved treatments, leaving substantial unmet medical needs for patients. Therefore, many of the considerations discussed above also apply to development programs for rare diseases.

FDA has a history of applying the philosophy underlying subpart E regulations to drugs for rare diseases. FDA recognizes that certain aspects of drug development that are feasible for common diseases may not be feasible for rare diseases and that development challenges are often greater with increasing rarity of the disease. The small population affected by a rare disease presents additional considerations that must be addressed and also calls for appropriate flexibility, discussed below.

1. Trial design

Because of the small number of patients with a rare disease, the number of patients eligible for enrollment in a trial may be small. In such situations, it is especially important to consider the advantages and disadvantages of various trial designs to achieve the objectives of establishing evidence of effectiveness as well as safety. Randomized, placebo-controlled trials with equal allocation are generally the most efficient designs to assess effectiveness; however, depending on the circumstances, sponsors should consider alternatives such as unequal allocation in a randomized controlled trial (i.e., more patients receive the new drug than the control), which can

[32] Section 526(a)(2) of the FD&C Act (21 U.S.C. 360bb(a)(2)).

provide increased safety experience and reduce the use of placebo, or a dose-comparison design (i.e., randomization to more than one dose, with or without placebo). If the effect of the drug can be discerned relatively quickly after starting or discontinuing the drug, designs such as cross-over trials, randomized withdrawal, or randomized delayed start should also be considered. Sometimes, as noted previously, a single-arm trial with an external control is an appropriate option. The ability of these or other trial designs to generate substantial evidence of effectiveness is dependent on the specifics of each situation.

Sponsors of drugs intended for rare diseases should consider designing their first-in-human trial to be an adequate and well-controlled clinical investigation that has the potential, depending on the trial results, to provide part of the substantial evidence of effectiveness to support a marketing application.[33]

2. Trial endpoints

Understanding of the pathophysiology of the underlying disease is important in planning clinical trials, including selection of endpoints. For many rare diseases, well-characterized clinical efficacy endpoints appropriate for the disease may need to be developed. In cases where utilizing clinical endpoints is not feasible because changes in symptoms and disease status occur too slowly to be measured in a clinical trial of reasonable duration, surrogate endpoints may be considered. It will be particularly important to understand the pathophysiology and natural history of the disease to help identify potential surrogate endpoints.

3. Number of trials

A second trial may be infeasible in certain rare disease settings where the limited patient populations preclude the conduct of a second trial. A similar situation may also arise when a drug is developed to target, for example, a low-frequency, molecularly defined subset of a more common disease and it may not be possible to screen and enroll enough patients within a reasonable period of time to conduct the second trial.[34] In these cases, the substantial evidence of effectiveness would typically be provided by a single trial plus confirmatory evidence.

4. Statistical considerations

As noted above, treatments for rare diseases often are intended to address unmet medical needs, and the considerations of balancing the harmful consequences of false positive and false negative results will often apply. In addition, the amount of evidence that can practically be acquired may be limited by the number of patients who can be recruited for trials. FDA may interpret the substantial evidence standard flexibly considering the harmful consequences of false negative and false positive results and the amount of evidence that can practically be acquired. Statistical approaches to evaluating treatments for rare diseases should consider the feasibility of trial

---

[33] Draft guidance for industry *Human Gene Therapy for Rare Diseases* (July 2018). When final, this guidance will represent the Agency's thinking on the topic is addresses.

[34] Guidance for industry *Developing Targeted Therapies in Low-Frequency Molecular Subsets of a Disease* (October 2018).

design, sample size, and endpoints, using methods and thresholds for demonstrating substantial evidence that are appropriate to these settings.

C. When conducting a human efficacy trial is not ethical or feasible

When it is not ethical or feasible to conduct clinical trials, FDA can allow the use of appropriate animal models to generate evidence to establish effectiveness for products intended to treat or prevent serious or life-threatening conditions caused by exposure to toxic biological, chemical, radiological, or nuclear substances. FDA's regulation governing these trials is known as the Animal Rule.[35]

[35] The Animal Rule "applies to certain new drug products that have been studied for their safety and efficacy in ameliorating or preventing serious or life-threatening conditions caused by exposure to lethal or permanently disabling toxic biological, chemical, radiological, or nuclear substances." 21 CFR 314.600; see also 21 CFR 601.90 (same restriction with respect to biological products).